

# POPULATION GENETIC ANALYSIS OF ENTIRE GENOMES, FROM SNP DISCOVERY TO GENOME-WIDE SCANS FOR SELECTION

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Mark Hamilton Wright

May 2010

© 2010 Mark Hamilton Wright  
ALL RIGHTS RESERVED

POPULATION GENETIC ANALYSIS OF ENTIRE GENOMES, FROM SNP  
DISCOVERY TO GENOME-WIDE SCANS FOR SELECTION

Mark Hamilton Wright, Ph.D.

Cornell University 2010

The analysis of molecular genetic data has driven the fields of molecular biology, genetics, population genetics, and quantitative genetics for over half a century. Only recently though has technology advanced to the point where molecular genetic data can be acquired cheaply and efficiently for the entire genome of several individuals enabling scientists to conduct genome-wide comparisons between several individuals or several population samples, and ask comprehensive questions regarding the nature of genetic variation in extant populations and the evolutionary forces in the population's history which generated and influenced this variation. Several challenges exist to utilizing these new technologies successfully however and in most cases both experimental optimization of laboratory protocols and the customization or *de novo* implementation of computational and statistical analysis methods are required to obtain adequate results. Even when the raw physical data acquired by these technologies has been successfully rendered into biologically meaningful molecular genetic data, the analysis of these large, genome-wide datasets is formidable and again requires advanced and customized methods to ask biologically motivated questions and produce conclusive results which may not have been obtainable without complete genome information. Here, I discuss two main technologies for the acquisition of genome-wide molecular data, next-generation sequencing technologies and fixed-array highly multiplexed SNP genotyping, and discuss the challenges in applying them in plant systems. Additionally, I demonstrate a population genetic analysis for the detection of recent selective sweeps in four subpopu-

lations of *Oryza sativa* (cultivated Asian rice) and one *Oryza rufipogon* population (wild Asian rice) utilizing the genome-wide molecular data acquired by next-generation sequencing. The development of an improved and accurate statistical method to detect selection in population genomic analysis combined with genome-wide data in each of these subpopulations allowed the extent and location of selective sweeps in *Oryza sativa* subpopulations and its wild progenitor *Oryza rufipogon* to be quantified and compared for the first time, revealing that each cultivated subpopulation appears to have a largely unique and independent selective and domestication history, but several advantageous alleles for cultivation of rice that originated and were selected for in one subpopulation have been introduced into other subpopulations by introgression.

## **BIOGRAPHICAL SKETCH**

Mark Wright was born in Hamilton, New York on July 20, 1976 to Sherry Stasse Wright and David Denio Wright, and spent the first 18 years of his life living in the nearby very small town of Waterville, New York before finally escaping. In August of 1994 at 18 years of age, Mark moved to Ithaca, New York to attend Cornell University as an undergraduate in the College of Arts and Sciences. Upon arriving at Cornell, Mark met several other students named Mark and decided to distinguish himself from others using his popular high school nickname “Koni.” The nickname “Koni” comes from “Markoni”, which itself refers to the Italian physicist Guglielmo Marconi whose famous experiment that lead to the development of the wireless telegraph and modern radio was accidentally repeated by Mark during an unauthorized little experiment with charging large storage batteries at very high voltages at Waterville High School. The high school’s aging PA system received and broadcasted the signal generated by the high voltage arcs during a regular announcement from the principal, leading to Mark getting caught and donned with the nickname “Markoni”, quickly popularized by both fellow students and faculty, and later shortened to just “Koni.” To this day, Mark is known to most only by the name Koni.

Largely undecided as a freshman regarding which field of study to pursue, Mark took classes in several subjects, initially intending to major in Chemistry, then Psychology, then Music, and finally deciding on Computer Science. Mark had an interest and aptitude in computers and programming from a young age, most likely due to early exposure to home computers of the 1980s such as the Commodore VIC-20 and 64 and the Apple II series, but before matriculating at Cornell Mark had no experience with IBM PC systems or the Macintosh systems which were replacing the early Apple II. In the summer of 1995, Mark found a summer job at Cornell working with Dr. Steven

Caldwell's research group in Sociology that was developing an extensive computer simulation of United States and Canadian populations for the purpose of analyzing different government policy proposals such as Social Security Reform and public health care policies. The position was originally advertised for programmers with comprehensive knowledge of the C programming language (which Mark did not have), and although the research group had filled their programming positions, they decided to hire Mark anyway to do some of the more laborious and tedious work that the programmers didn't have time for, which was mainly producing quality graphs and tables from the simulation model's extensive output. The research group used PC computers and an operating system called "OS/2" which Mark had never heard of. After some quick learning, Mark again demonstrated unusual aptitude in programming computers, by developing a way to programmatically generate the more than 10,000 graphs in an automated fashion in less than a hour which originally Dr. Caldwell and the research team expected Mark to make by hand one at a time and take the entire summer or longer to complete. After the summer of 1995, Mark decided to pursue Computer Science as a major and received A or A+ marks in nearly all computer science classes. In addition to completing the 4-year curriculum of the Computer Science major in 3 years (having started in his sophomore year), Mark also took a number of graduate level classes in Computer Science and specialized largely in Distributed Systems with an interest in security and cryptography protocols required for distributed computing in an untrusted environment or network. Annoyingly, Mark graduated from Cornell in May 1998 but he was certainly not finished with his studies.

Aside from a 3 month trip to Spain where Mark attempted to learn a foreign language and experience a different culture, Mark remained in Ithaca, New York and upon returning from Spain Mark resumed work with Dr. Steven Caldwell as a programmer,

continuing to develop the microsimulation model Mark worked on as a student. During this time as a professional programmer, Mark mastered many advanced programming techniques that have served him well in his later pursuit of population genetics and computational biology.

After approximately 3 years working with Dr. Caldwell but at a small company outside of Cornell University, Mark was eager to obtain a position at Cornell in order to have the benefit of being able to take classes again. A good friend, Dan Ilut, had recently taken a job with Dr. Steven Tanksley in the Department of Plant Breeding and Genetics, and while Dan's initial work exceeded Dr. Tanksley's expectations, Dr. Tanksley was eager to include more computational work in his research and opened a second position which Mark applied for and was successful. At this time, Mark had little to no knowledge of biology or genetics, and initially learning was slow. After two years Mark applied to four graduate schools for a Ph.D. program in Computer Science, and was rejected at all four schools despite perfect exam scores and excellent transcripts. In the time between application and the ultimate decisions however, Mark experienced a transition of interests from distributed systems and cryptography, to biology and genetics which he was increasingly exposed to working with Dr. Tanksley's group.

With the support and encouragement of Dr. Tanksley, Mark decided to apply for graduate school at Cornell in Computational Biology and Biological Statistics (then Biometry) in the employee degree program (EDP). Mark's application was accepted, but no funding was promised and it was understood that Mark's tuition would be covered by the EDP benefits. The EDP program only allows one class at most per semester and this proved quite frustrating for Mark who made slow progress with coursework over the next two years and coming from a primarily computer science background,

Mark required more coursework to address deficiencies in both Biology and Statistics. Mark's adviser in Biometry, Rasmus Nielsen, decided to leave Cornell before Mark could complete enough coursework and gain enough background knowledge to begin research with him.

Without an adviser in Biometry, an increasing interest in experimental biology, and frustration with the slow progress of the EDP, Mark decided to re-apply to graduate school for full time study and full time funding in the field of Plant Biology. Mark's application was successful and Mark began full time study, largely starting over, in Plant Biology in August of 2004. Initially Mark was eager to learn molecular biology and experimental techniques, and was quickly exposed to many through laboratory classes and lab rotations with different groups. During this first year, Mark met his girlfriend of the last 6 years, Elhan Ersoz, who stirred interest in population genetics and quantitative genetics. During the course of first year rotations, Mark elected to rotate outside the field of Plant Biology with Dr. Charles Aquadro whose research primarily involved experimental population genetics. Mark's interest in population genetics grew substantially due to the influence of both Dr. Aquadro and Elhan, ultimate resulting in Mark electing to change fields of study again from Plant Biology to Genetics and Development.

For the next two years of graduate school Mark worked with Dr. Aquadro's group and completed all necessary coursework and teaching requirements, but struggled to develop a dissertation research agenda. Due to an unfortunate struggle with severe mental illness, Mark strongly considered abandoning his graduate program at the end of his 3rd year and return to full time programming which Mark felt he could continue to perform well despite a chronic illness which was not responding well to treatment. At this time, Mark approached Dr. Carlos Bustamante, a member of Mark's graduate commit-



tee, about the possibility of going on a leave-of-absence from the graduate school and working with his group as a programmer rather than a graduate student. Dr. Bustamante strongly advised not leaving graduate school and suggested that Mark try working with his group doing some programming-intensive analyses but remain a graduate student.

This proved a successful combination as the larger, genome-wide datasets available in the Bustamante Lab were well matched to Mark's by far strongest skill: designing and programming efficient, fast, and parallelized algorithms. Through Dr. Bustamante's existing collaborations on campus, Mark was introduced to Dr. Susan McCouch of Plant Breeding and Genetics, and ultimately Mark would work primarily with her on the majority of the research presented in this dissertation. A long term goal of Mark's graduate school experience is to learn and develop the critical thinking skills of a scientist, a necessary skill to follow the efficient analysis of genome-wide large datasets in order to produce publishable results. Mark feels these goals have largely been obtain, and is ready to proceed to the next stage in his career. Following graduation, Mark intends to further pursue much of the research presented here together with Dr. Susan McCouch.

This document is dedicated to Elhan Ersoz, who has inspired so much and endured even more.

## ACKNOWLEDGEMENTS

I would like to acknowledge my committee for their support and complementary influence on my thinking and academic development throughout graduate school. **Dr. Carlos Bustamante**, my committee chair, has financially supported my work as a graduate student for the past 3 years and has taught me statistics and the application and development of population genetics in the context of whole-genome datasets. Dr. Bustamante also took me as struggling graduate student starting in my 4th year and helped me bridge the gap between my computational skills and biology through access to many population genomic datasets that required advanced computational methods. **Dr. Edward Buckler** has taught me the meaning of “thinking big” in experiments and this is so critical in whole genome studies. **Dr. Charles Aquadro** supported me financially in my 3rd year of graduate school and graciously took me on as a student before I ultimately moved to the Bustamante lab. Through my time in Dr. Aquadro’s lab I learned much about experimental population genetics and acquired the ability to think in a more hypothesis driven fashion. Since many population genomic studies today are more data driven than hypothesis driven, I feel Dr. Aquadro’s contribution to my scientific development and influence was critical in becoming a good scientist today.

Next, I would like to thank **Dr. Susan McCouch** who I have worked with extensively and should be considered an unofficial member of my committee. Projects with Dr. McCouch have been the source of much of my successful work as a graduate student and encompass two of the three chapters presented here. I am indebted to her for these opportunities and for the enthusiasm she brings to her work and her life which inspired me to work hard on these projects.

I am also indebted to my parents **David and Sherry Wright** for their continued support in academic pursuits. Without their support, both morally and financially, I could not have pursued education to this level or had the opportunity to develop intellectually

so far.

I would like to thank **Dr. Robert Mendola, MD, Thomas Cullen,** and **Dr. Gregory Eells** for their help in a very difficult time and continued support and care over the past 3 years. Also, I would like to thank **Dr. Charles Aquadro** of my committee, who was my adviser when severe mental illness struck me leading to me to seek the help of the Dr. Mendola, Tom Cullen, and Dr. Eells for treatment. Dr. Aquadro was very understanding and patient at these times, and had the attention and presence of mind to point out a change in my thinking, behavior, and speaking that was ultimately key to a correct diagnosis and effective treatment. Many other advisers would have neither the time nor the interest to deal with a student so distressed let alone make a critical contribution to their care and recovery. As much as I am ashamed of the madness in myself that I exposed to Dr. Aquadro, I am very grateful for his support and his help.

**Andy Reynolds**, a programmer in the Bustamante Lab, has worked with me most closely many of these projects and has been a good friend. **Chih-Wei Tung**, post-doc in the McCouch Lab, performed most of the molecular biology and laboratory work that generated the raw data for the methods I develop in this dissertation and use for population genetic analyses.

For the selective sweep analysis in rice, most data was generated in the McCouch lab but data was also contributed by **Dr. Masahiro Yano** of Japan, **Dr. Yue-Fe Hsing** of Taiwan, **Dr. Brian Scheffler** of the USDA and **Dr. Adam Price** of the United Kingdom.

I am indebted to my friend and colleague **Dan Ilut** who first introduced me to bioinformatics and computational biology and got me started back in 2001 in a new job at Cornell in this field with **Dr. Steven Tanksley** of the Plant Breeding department. Dr. Tanksley greatly encouraged my interest in biology and plant genetics and has been a tremendous positive influence over my career in science.

I would like to thank **Jennifer Vorhoff** for stimulating company and conversation during the frantic time of writing and compiling this dissertation.

And finally, I would like to thank my girlfriend **Dr. Elhan Ersoz** for her support and faith in me during my graduate school career. Elhan initially inspired me to pursue population genetics and continues to inspire me in this direction and others today. She also has been a wonderful friend and a loyal companion.

## TABLE OF CONTENTS

|   |           |
|---|-----------|
| Biographical Sketch . . . . .   | iii       |
| Dedication . . . . .  | viii      |
| Acknowledgements . . . . .  | ix        |
| Table of Contents . . . . .   | xii       |
| List of Tables . . . . .  | xiii      |
| List of Figures . . . . .   | xiv       |
| <b>1 Introduction</b>   | <b>1</b>  |
| <b>2 Large-scale Enrichment and Discovery of Gene-enriched SNPs in Maize</b>  | <b>7</b>  |
| 2.1 Introduction . . . . .  | 7         |
| 2.2 Materials And Methods . . . . .   | 10        |
| 2.3 Results . . . . .   | 19        |
| 2.4 Discussion . . . . .  | 31        |
| <b>3 ALCHEMY: A Reliable Method for Automated SNP Genotype Calling for Small Batch Sizes and Highly Homozygous Populations</b>  | <b>38</b> |
| 3.1 Introduction . . . . .  | 39        |
| 3.2 Approach . . . . .  | 42        |
| 3.3 Algorithm . . . . .   | 46        |
| 3.4 Methods . . . . .   | 49        |
| 3.5 Results . . . . .   | 52        |
| 3.6 Discussion . . . . .  | 57        |
| <b>4 Robust Composite Likelihood Method for Genome Wide Selection Scans Reveals a Largely Independent Selective History of <i>Oryza sativa</i> Subpopulations Combined with Introgressions of Selected Alleles Between Subpopulations</b> | <b>62</b> |
| 4.1 Introduction . . . . .  | 63        |
| 4.2 Methods . . . . .   | 69        |
| 4.3 Results . . . . .   | 75        |
| 4.4 Discussion . . . . .  | 92        |

## LIST OF TABLES

|     |   |    |
|-----|---|----|
| 2.1 | Screening and reference databases used . . . . .  | 14 |
| 2.2 | Sequence composition of modified HMPR and UF libraries . . . . .  | 22 |
| 2.3 | Gene Enrichment Analysis of modified HMPR and UR libraries . . . . .  | 24 |
| 2.4 | Summary of the assembly process . . . . .   | 25 |
| 2.5 | Summary of putative SNP calls with and without PDL . . . . .  | 30 |
| 2.6 | Summary of B73/Mo17 454 SNP validation . . . . .  | 31 |
| 3.1 | Comparison of reference line genotype calls to published genome se-<br>quence . . . . .                         | 54 |
| 3.2 | Pairwise concordance in genotype calls for replicate samples . . . . .  | 55 |
| 3.3 | ALCHEMY vs. BRLMM-P on human Hapmap samples . . . . .   | 56 |
| 3.4 | ALCHEMY vs. BRLMM-P on small sample batches . . . . .   | 57 |
| 4.1 | Number of sweeps called in <i>O. sativa</i> subpopulations and <i>O. rufipogon</i> . . . . .                    | 84 |
| 4.2 | Haplotype sharing between subpopulations at subpopulation-specific<br>selective sweep loci . . . . .            | 88 |
| 4.3 | Haplotype sharing between subpopulations at loci where both subpop-<br>ulations show selective sweeps . . . . . | 91 |

## LIST OF FIGURES

|     |  |    |
|-----|--|----|
| 2.1 | Schematic of the PDL method . . . . .  | 27 |
| 3.1 | Density plot of probe intensity distribution . . . . .   | 43 |
| 3.2 | Effect of increasing number of samples simultaneously analyzed . . . .                           | 58 |
| 4.1 | Effect of demography and recombination rate on CLR null distribution                             | 77 |
| 4.2 | QQ-plot of expected and observed p-values calculated by the permuta-<br>tion procedure . . . . . | 80 |
| 4.3 | Typical CLR distribution and its prediction by log-linear regression . .                         | 82 |
| 4.4 | Selective sweep map of <i>O. sativa</i> and <i>O. rufipogon</i> . . . . .                        | 85 |



## CHAPTER 1

### INTRODUCTION

Prior to the discovery of the molecular structure of deoxyribonucleic acid (DNA) (Watson and Crick, 1953), or the fact that this molecule carried hereditary information, an impressive amount of insight into genetics was achieved starting at the turn of the previous century with the rediscovery of Mendel's seminal work 40 years earlier and concept of genetic linkage, first demonstrated by Morgan, Sturtevant, Muller, and Bridges in 1915 (Morgan *et al.*, 1915). Just prior to the publication of this seminal treatise on the mechanisms of Mendelian inheritance, Sturtevant had published the first genetic linkage map in 1913, deducing the linear ordering and approximate distances between six sex-linked genes in *Drosophila melanogaster* (Sturtevant, 1913). This first genetic map and many to follow were created using only phenotypic observation and inferring the underlying genetics and with no knowledge of the underlying physical nature of molecules which were responsible for Mendelian transmission of heritable characters from parents to progeny.

Following the discovery of the double-helix structure of DNA and its long linear organization in eukaryotes, the discovery of restriction enzymes (Meselson and Yuan, 1968) led to the birth of the molecular genetics, or the ability to directly assay the hereditary material itself as opposed to requiring a phenotypic realization and inheritance pattern to discern the genetic state. Using restriction enzymes, the first molecular genetic "markers" such as restriction fragment length polymorphism (RFLP) markers were developed and provided researchers with enough heritable and polymorphic loci across the genome to construct the first complete genetic linkage map, spanning the entire genome and enabling for the first time the resolution of complex quantitative traits to Mendelian

factors (Paterson *et al.*, 1988) With the development of the polymerase chain reaction (PCR) (Mullis and Faloona, 1987), it became possible to cheaply and easily assay a new type of molecular marker called simple sequence repeats (SSR) or microsatellites. With molecular genetic maps becoming more and more densely populated with closely spaced markers across the genome, and increasing ease and decreasing cost of assaying large populations, it became possible to use these markers to study the nature of genetic variation in populations, to associate molecular markers with phenotypes of interest, and to use these associations to assist the development of new crop varieties and improve animal breeding practices.

The ability to obtain complete sequence of the DNA molecule represented another leap forward in molecular genetics, for the first time completely revealing the underlying molecule that is responsible for nearly all inheritance from parents to their progeny and the ultimate source of what precisely defines the genetic differences in individuals of any population. While early sequencing was a manual and laborious procedure, advances in technology and engineering adapted the basic cycle-sequencing or Sanger sequencing method to an automated scale, eventually resulting in the complete genomic sequence for several model systems (Blattner *et al.*, 1997; Zagulski *et al.*, 1998; Adams *et al.*, 2000; Arabidopsis Genome Initiative, 2000; Goff *et al.*, 2002; Yu *et al.*, 2002), and most notably, the complete Human genome (Lander *et al.*, 2001; Venter *et al.*, 2001). Even before the complete genome sequence of a single individual from any of these organisms were assembled, researchers in population genetics began sequencing specific loci in several individuals and comparing these sequences to learn about the extent and frequency of genetic polymorphisms in natural populations and to develop and refine models of how these polymorphisms came into existence (eg, mutation), how and whether they proliferated in the population, and how they are either lost or eventually

fixed, ceasing to be polymorphic.

Sequencing of several individuals, now known as “resequencing” as it is often done with the use of a complete fully assembled genome sequence of one individual as a reference, quickly revealed that the most common molecular polymorphism observed in sequence data is that of a single base substitution between two individuals, commonly known as a SNP (single nucleotide polymorphism). Conveniently, it is easy to develop these molecular markers into assays which directly determine the state of the polymorphic nucleotide, without the need to sequence the entire region for each individual. Initially, adoption of SNP-based molecular assays was slow due to the fact that their bi-allelic nature, despite their ease of discovery, is less informative per molecular genotype than SSR markers which typically harbor multiple alleles and at appreciable frequencies in the population and therefore are more likely to differentiate individuals.

This has changed recently due to the development of high-throughput, highly multiplexed fixed array SNP detection assays, with commercial products ranging from 384 simultaneous SNP assays per sample to over 1 million SNPs per sample. With SNPs being present nearly everywhere in the genome and at a much higher density than SSRs, the lower information per genotype of SNPs compared to SSRs is more than compensated for by the far greater number of SNPs that can be assayed at equivalent costs. Only recently now, this has enabled for the first time the possibility of densely genotyping a large population sample, with applications in complex trait dissection, heritable disease genetics, as well as large scale population genetic studies to estimate the recent history of a population, its ancestral origin, structuring within and between populations, and the extent to which Darwinian selection may have influenced the genetics of extant popula-

tions.

Initially, the ability to cheaply assay SNPs in these highly multiplexed formats preceded the ability to perform an adequate genome-wide discovery and cataloging of SNPs. The complete *de novo* sequencing of a single individual using automated Sanger sequencing was and still is a very time consuming and costly process. During the course of the research presented here, a major transformation in DNA sequencing technology took place, first with the introduction of the 454 GS FLX which produced up to 100 Mb of DNA sequence per run at a fraction of the time and cost for Sanger sequencing. This technology was quickly followed by the Illumina GenomeAnalyzer which initially provided 1000 Mb of sequence per run and recent upgrades of this system in just the last year are now producing up to 30 Gb of sequence per run, at a fraction of the cost of 454 sequencing. While the read lengths produced by these technologies is often not sufficient to perform a *de novo* assembly of a genome, the existence of a fully assembled and complete reference genome allows the millions of short reads produced by these technologies to be uniquely placed against the reference in a majority of cases, revealing polymorphisms between the *re*sequenced individuals and the reference. The development of these technologies has resulted in cheap and efficient SNP discovery for the purposes of developing fixed-array SNP genotyping products and even now these methods are enabling the direct resequencing of large collections of samples for population genetic analysis.

Although these technological advancements have presented tremendous opportunity, using them in any particular system of study requires optimization of experimental protocols and development of custom or tailored analysis tools in order to achieve the de-

sired results. Presented in this dissertation, I discuss the use of 454 resequencing for SNP discovery in *Zea mays* (maize) and the challenges presented by the complexity of the maize genome and how they were solved by a custom and novel analysis methodology. Following, two different fixed-array SNP genotyping products were developed in *Oryza sativa* (domestic rice) for purpose of genotyping an association population consisting of cultivated and wild (*O. rufipogon*) rice varieties. In both cases, the technologies marketed by the vendor were originally designed and tested for applications in human populations, but our largely inbred rice population differed significantly from human populations in that heterozygote genotypes were rare. In order to obtain accurate genotype calls in rice, a new analysis method called “ALCHEMY” was developed to estimate and incorporate inbreeding levels into a statistical model of the underlying assay’s raw data. This resulted in a great improvement in accuracy and call rates over the vendor’s software and may have applications in other systems where individuals are largely homozygous within sample batches.

Finally, using the Illumina GenomeAnalyzer IIx platform, we obtained resequencing data for the entire genome of 39 cultivated rice varieties representing 4 major subpopulations, and 8 *Oryza rufipogon* varieties. Initially this resequencing effort was aimed solely at improving SNP discovery in rice for the design of a larger and comprehensive fixed-array genotyping product. Here, I use the SNP discovery data obtained so far in the on-going SNP discovery effort to conduct a genome-wide scan for recent strong selection in the rice genome, for each subpopulation and the wild progenitor *O. rufipogon*. In order for an adequate comparison of the extent and locations of selective events in rice genome between subpopulations, I extend a composite-likelihood method of Nielsen *et al.* (2005) with an extensive permutation test procedure and show that the inference of selection is robust to non-selective factors such as population size changes,

varying mutation rates and SNP densities, and varying recombination rates. All previous genome-wide selection scan methods employed to date are plagued by these problems and force researchers to guard against false positives by using strongly conservative assumptions and parameters. The more accurate statistical assessment of selection presented here allows for a more meaningful comparison between subpopulations and their selective histories, and suggests exciting further approaches to reveal how these subpopulations were domesticated, where domestication traits originated, and how they were transferred among subpopulations.

## CHAPTER 2

# LARGE-SCALE ENRICHMENT AND DISCOVERY OF GENE-ENRICHED SNPS IN MAIZE

Published in January 2009 as a co-first-author paper with Michael Gore: M.A. Gore, M.H. Wright, E. S. Ersoz, P. Bouffard, E.S. Szekeres, T.P. Jarvie, B.L. Hurwitz, A. Narechania, G.S. Grills, D.H. Ware, E.S. Buckler. *Large-scale enrichment and discovery of gene-enriched SNPs*. The Plant Genome (2009) 2:121-133. Used under E. Buckler's license as a USDA employee

**Abstract.** Whole-genome association studies of complex traits in higher eukaryotes require a high density of single nucleotide polymorphism (SNP) markers at genome-wide coverage. To design high-throughput, multiplexed SNP genotyping assays, researchers must first discover large numbers of SNPs by extensively resequencing multiple individuals or lines. For SNP discovery approaches using short read lengths that next-generation DNA sequencing technologies offer, the highly repetitive and duplicated nature of large plant genomes presents additional challenges. Here, we describe a genomic library construction procedure that facilitates pyrosequencing of genic and low-copy regions in plant genomes, and a customized computational pipeline to analyze and assemble short reads (100-200 bp), identify allelic reference sequence comparisons, and call SNPs with a high degree of accuracy. With maize (*Zea mays* L.) as the test organism in a pilot experiment, the implementation of these methods resulted in the identification of 126,683 putative SNPs between two maize inbred lines at an estimated false discovery rate (FDR) of 15.1%. We estimated rates of false SNP discovery using an internal control, and we validated these FDR rates with an external SNP dataset that was generated using locus specific PCR amplification and Sanger sequencing. These results show that this approach has wide applicability for efficiently and accurately detecting gene-enriched SNPs in large, complex plant genomes.

## 2.1 Introduction

The average nucleotide diversity of coding regions between any two maize lines ( $\pi=1-1.4\%$ ) is 2- to 5-fold higher than other domesticated grass crops (Buckler *et al.*, 2001;

Tenaillon *et al.*, 2001; Wright *et al.*, 2005). Moreover, it is not uncommon to find maize haplotypes more than 2% diverged from one another (Tenaillon *et al.*, 2001; Wright *et al.*, 2005) and even as high as 5% (Henry and Damerval, 1997). Intragenic linkage disequilibrium (LD) rates rapidly decline to nominal levels within 2 kb in a population of diverse maize inbred lines (Remington *et al.*, 2001). Of the  $\approx 2500$  Mb that constitutes the maize genome, less than 25% is genic or low-copy-number sequence, with large blocks of highly repetitive DNA such as retrotransposons intermixed throughout (Hake and Walbot, 1980; Meyers *et al.*, 2001; SanMiguel *et al.*, 1996). Retrotransposons are generally recombinationally inert, and most meiotic recombination in the maize genome is restricted to gene-rich regions (Fu *et al.*, 2002, 2001; Yao *et al.*, 2002). Association mapping strategies, which rely on ancient recombination for dissecting complex traits, require that SNPs within these recombinationally active gene regions be identified and genotyped in phenotypically diverse populations (Zhu *et al.*, 2008). Because of the rapid decay of intragenic LD in a highly diverse genome with an estimated 59,000 genes (Messing *et al.*, 2004), several million gene-enriched SNP markers may be necessary for whole-genome association studies in diverse maize (E. Buckler, unpublished).

Retrotransposons contain a higher density of methylation in the form of 5-methylcytosine relative to genic sequences – a property unique to plant genomes (Rabinowicz, 2003; Rabinowicz *et al.*, 2005). HypoMethylated Partial Restriction (HMPR) is a library construction method that exploits this property to facilitate the efficient sequencing of gene rich regions in large, highly repetitive plant genomes (Emberton *et al.*, 2005). The principle underlying HMPR is that the complete digestion of plant genomic DNA with a 5-methylcytosine-sensitive (MCS) restriction enzyme that has a 4 bp recognition sequence permits the fractionation of genic and repetitive DNA by gel electrophoresis. Large restriction fragments (20-150 kb) contain blocks of highly



methyated retrotransposons, while much smaller fragments (<1000) comprise a fraction that is gene-enriched (Bennetzen *et al.*, 1994; Yuan *et al.*, 2002). Emberton *et al.* (2005) used a partial digestion of maize genomic DNA with a MCS 4 bp cutter, followed by gel-purification and cloning procedures to construct maize HMPR libraries that contained larger (1-4 kb), overlapping gene fragments more suitable for Sanger sequencing read lengths (800-1200 bases). These maize HMPR libraries showed more than 6-fold enrichment for genes compared to control libraries. This level of gene enrichment was comparable to that achieved by other non-transcriptome-based gene-enrichment sequencing technologies tested on maize (Gore *et al.*, 2007; Palmer *et al.*, 2003; Rabinowicz *et al.*, 1999; Whitelaw *et al.*, 2003; Yuan *et al.*, 2003), but maize HMPR libraries were superior for repeat elimination and enrichment of low-copy, non-coding sequences.

With the recent emergence of “next-generation” DNA sequencing technologies it is technically feasible to economically and rapidly resequence hundreds of millions of bases. Using these high-throughput sequencing-by-synthesis (Margulies *et al.*, 2005) or sequencing-by-ligation (Shendure *et al.*, 2005) technologies in a read-to-reference based SNP discovery approach presents computational challenges because the length and quality of obtained individual reads are shorter and potentially of lower fidelity than single-pass Sanger sequencing reads. Furthermore, the maize genome is the product of ancient and perhaps more recent tetraploidization and rearrangement events (Gaut and Doebley, 1997; Swigonova *et al.*, 2004; Wei *et al.*, 2007), and as a result contains a high proportion of duplicated genes (Blanc and Wolfe, 2004; Emrich *et al.*, 2007; Messing *et al.*, 2004). This confounds the unique mapping of short reads if duplicated genes (i.e., paralogs) are recently diverged and thus nearly identical in nucleotide sequence. Recently, a computational SNP calling pipeline built on the POLYBAYES polymor-

phism detection software (Marth *et al.*, 1999) and “monoallelism” rules was developed and used to analyze expressed sequence tags (ESTs) that were obtained by 454 pyrosequencing of cDNAs prepared from two maize inbred lines (Barbazuk *et al.*, 2007). This pipeline reduced the number of false positive SNPs that resulted from sequencing errors and alignment of paralogous sequences, which facilitated the identification of more than 7,000 putative SNPs in expressed genes.

Nonetheless, if the discovery of maize SNP markers on the order of millions is to be economically viable, the use of low cost, next-generation DNA sequencing technologies is clearly required. These high-throughput DNA sequencing technologies can be more efficiently used in the large-scale discovery of SNPs for maize association mapping studies if resequencing is concentrated within the recombinationally active gene regions of the vastly repetitive maize genome. The objectives of this study were (i) to adapt HMPR gene-enrichment sequencing to a massively parallel pyrosequencing platform and (ii) to develop a read-to-reference based SNP calling pipeline for short reads (100-200 bp) that maximizes SNP detection power, while controlling the number of detected false positive SNPs resulting from sequencing errors and the alignment of paralogous sequences.

## **2.2 Materials And Methods**

**DNA Isolation from Maize.** We extracted nuclear DNA from nuclei prepared from etiolated (pale green), inner husk leaves (100 g) of field-grown maize inbred line B73 as previously described by Rabinowicz (2003). A more specialized cultivation technique was required to obtain genomic DNA from maize root tissue. Kernels from maize inbred

lines B73 and Mo17 were surface sterilized in a 10% (vol/vol) bleach solution (5.25% Sodium Hypochlorite) by gently rocking for 30 min, followed by 3X 10 min rinses with sterile water. The kernels were left to imbibe overnight in sterile water at room temperature with gentle rocking. Ten kernels were placed in a vertically orientated seed germination pouch (Mega International, West St. Paul, MN) and germinated in a dark growth chamber held at 28 C. Roots of 1-wk-old maize seedlings were bulk harvested and immediately frozen in liquid N<sub>2</sub> prior to storage at -80 C. Total genomic DNA was isolated from homogenized frozen 1-week-old root tissue using the DNeasy Plant Maxi Kit (QIAGEN, Valencia, CA) according to the manufacturer's protocol.

**Modified HMPR Library Construction.** Complete digestions of 5  $\mu$ g of maize husk nuclear DNA (B73) and seedling root total genomic DNA (B73 and Mo17) were individually performed in 100  $\mu$ L volumes with 50 U of HpaII (New England Biolabs, Ipswich, MA) at 37 C for 16 h, followed by heat inactivation of the enzyme at 65 C for 20 min. HpaII fragments ranging in size from >10 kb to less than 100 bp (data not shown) were separated on a low melting 0.8% SeaPlaque agarose gel (Cambrex Bio Science Rockland, Inc., Rockland, ME). Restriction fragments ranging in size from 100-600 bp were excised from the gel and purified using the QIAquick Gel Extraction kit (QIAGEN, Valencia, CA), according to the manufacturer's protocol. Gel-isolated HpaII fragments were randomly ligated to each other with 1  $\mu$ L of highly concentrated T4 DNA ligase (20 U/ $\mu$ L) (New England Biolabs, Ipswich, MA) in a total reaction volume of 20  $\mu$ L at 16 C for 16 h, followed by heat inactivation of the enzyme at 65 C for 20 min.

Several micrograms of concatenated HpaII fragments were needed for the downstream nebulization procedure (see 454 sequencing and data processing section). How-

ever, this would typically require low-throughput, large-scale DNA extractions and gel isolations, because an estimated 95% of the maize genome was intentionally discarded. Alternatively, we found it more efficient to generate microgram quantities of concatenated HpaII fragments using Phi29-based isothermal amplification of long concatemer templates in a nanogram-scale reaction. Briefly, the GenomiPhi V2 DNA Amplification Kit (GE Healthcare, Piscataway, NJ) was used to amplify 1  $\mu$ L of the 10 ng/ $\mu$ L ligation reaction per the manufacturer's instructions. This kit uses the high fidelity Phi29 ( $\phi$ 29) DNA polymerase, dNTPs, and random hexamers to replicate linear genomic DNA by multiple displacement amplification. Several independent GenomiPhi amplification reactions were performed and pooled for each library to ensure a low level of amplification-induced bias. The GenomiPhi reaction was separated on a low melting 0.8% SeaPlaque Agarose gel, and amplification products ranging in size from 3-10 kb were isolated from the gel with the QIAquick Gel Extraction kit and used in the downstream 454 sample preparation procedure.

**454 Sequencing and Data Processing.** Sequence sample preparation and data generation were performed with the Phi29 amplified HpaII concatemer DNA of two B73 HMPR libraries (husk and root) and one Mo17 HMPR library (root) using the 454 GS FLX platform at 454 Life Sciences (Branford, CT). In addition, total genomic DNA isolated from the same seedling root tissue of B73 was sequenced on the same 454 platform, which served as an unfiltered (UF) genomic control to assess the level of gene-enrichment in modified HMPR libraries. Approximately 5  $\mu$ g of high molecular weight DNA was fragmented by nebulization to a size range of 300-500 bp. Preparation of 454 libraries, emulsion-based clonal amplification, library sequencing on the Genome Sequencer FLX System as well as signal processing and data analysis were performed as previously described by Margulies et al. (2005). Also, the 454 base-calling software

(version 1.1.03.24) provided error estimates (Q values) for each base, none of which exceeded a value of 40.

The expected yield per run of the 454 GS FLX is approximately 100 Mb, potentially more under ideal conditions. However, sequencing the B73 husk library with a single instrument run produced only 65.6 Mb of sequence because a less than optimal DNA copy per bead ratio was used for emulsion PCR. A more optimal DNA copy per bead ratio was used for the B73 root library, improving sequence yield to 101.3 Mb in a single run. The Mo17 root library was sequenced with four runs that in total yielded 236.7 Mb of sequence. This total sequence yield for the Mo17 root library was 41% lower than expected, indicating that further optimization was still needed. In addition, we sequenced (1 run; 130.9 Mb) randomly sheared B73 total genomic DNA, which served as the UF library. The raw 454 sequencing data are available in the NCBI Short Read Archive (<http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi>).

**Screening and Filtering of 454 Sequences.** Because modified HMPR libraries contained HpaII concatemers, 454 reads generated from sequencing these libraries were digested *in silico* at HpaII recognition sites (5'-C/CGG-3'). This was done to produce independent, non-chimeric HpaII fragment sequences. All 454 reads from the UF control library and HpaII fragment sequences less than 40 bp in length were discarded. HpaII fragment sequences and UF sequences (> 40 bp) were searched using BLAT (Kent, 2002) against The Institute for Genomic Research (TIGR) maize repeat database Version 4.0 ([http://maize.tigr.org/repeat\\_db.shtml](http://maize.tigr.org/repeat_db.shtml)) to identify repetitive sequences. Also, sequences were searched against mitochondrial (GenBank accession no. NC\_007982.1) and chloroplast (GenBank accession no. NC\_001666.2) genome sequences of maize. We performed BLAT searches with default parameters, except for

a tile size of 16. We considered BLAT similarities significant if the expectation value was less than  $10^{-5}$  and the local alignment length was 40 bp or longer. Sequences that had a significant match to a repeat sequence or an organellar genome were discarded. Remaining sequences were similarly searched with BLAT against the Maize Assembled Genome Island Version 4.0 Contigs and Singletons (MAGIv4.0 C&G) database (<http://magi.plantgenomics.iastate.edu/>). Because a large number of sequences did not match any sequences in the MAGIv4.0 C&G database, these unmatched HpaII fragment and UF sequences were also searched against the complete genome sequences of japonica rice (*Oryza sativa* L.) (<http://rice.plantbiology.msu.edu/>) and sorghum (*Sorghum bicolor* L.) (<http://www.phytozome.net/>) as well as maize expressed sequence tag (EST) sequences within the Dana-Farber Cancer Institute (DFCI) maize gene index release 17.0 (<http://compbio.dfci.harvard.edu/tgi/>). Sequences that did not have a significant match in any of these additionally searched databases were considered contaminant (non-maize) sequence and discarded. Summary statistics and source information for all databases are found in Table 2.1.

Table 2.1: Databases used for filtering and screening raw 454 sequence reads and the database used as reference sequence for calling SNPs.

| Name                            | Source  | Sequences | Mb    |
|---------------------------------|---|-----------|-------|
| Screening Databases             |   |           |       |
| Maize Chloroplast Genome        | GenBank acc. no. NC_001666.2  | 1         | 0.14  |
| Maize Mitochondrial Genome      | GenBank acc. no. NC_007982.1  | 1         | 0.569 |
| TIGR Maize Repeatv4.0           | <a href="http://maize.tigr.org/repeat.db.shtml">http://maize.tigr.org/repeat.db.shtml</a>   | 26,791    | 19.6  |
| Rice                            | <a href="http://rice.plantbiology.msu.edu/">http://rice.plantbiology.msu.edu/</a>           | 12        | 372.1 |
| Sorghum                         | <a href="http://www.phytozome.net/">http://www.phytozome.net/</a>                           | 10        | 697.6 |
| DFCI Maize Gene Index r17.0     | <a href="http://compbio.dfci.harvard.edu/tgi/">http://compbio.dfci.harvard.edu/tgi/</a>     | 115,744   | 86.3  |
| Reference Database              |   |           |       |
| MAGIv4.0 Contigs and Singletons | <a href="http://magi.plantgenomics.iastate.edu/">http://magi.plantgenomics.iastate.edu/</a> | 727,781   | 675.2 |

**Assembly of 454 Sequences.** We assembled the retained non-repeat HpaII fragment sequences into multiple sequence alignments using the CAP3 sequence assembly pro-

gram (Huang and Madan, 1999). The following CAP3 assembly options were used: -p 99 (overlaps must be  $\geq 99\%$  identity), -s 401 (alignment score must be  $>400$ , minimum value allowed), -h 3 (maximum overhang of 3%), and alignment scoring options (-m 20, -n 40, and -g 21) that allowed a perfect match overlap of 40 bp to satisfy the minimum alignment score for assembly. Additionally, CAP3 computed a Q value for each base of the consensus sequence. Assemblies were performed separately for B73 (husk and root) and Mo17 (root) non-repeat HpaII fragment sequences. We did not assemble UF sequences, as they were only used to measure the level of gene-enrichment and repeat depletion in modified HMPR libraries.

Because CAP3 could not execute with all sequences input at once, we performed a preliminary clustering of sequences into a collection of disjoint groups with no inter-group homology. Clustering was performed by a custom program in a manner equivalent to NCBI BLASTClust (available at <http://www.ncbi.nlm.nih.gov/BLAST/docs/blastclust.html>). We did not use BLASTClust because it could not run on our systems with the amount of input data supplied. CAP3 was then executed on each cluster separately. The preliminary clustering revealed that about 5% of sequences were still chimeric because of an HpaII site that was eliminated by a sequencing error or erroneous end-joining ligation. A simple modification to the clustering algorithm allowed almost all chimeras to be detected and split before CAP3 assembly.

We developed a custom program to analyze the CAP3 assembly output and extract a consensus sequence and associated CAP3-based Q values from each multiple sequence assembly as well as the number of sequences concordant with each consensus base

(coverage depth). Because of partial overlaps and potential disagreements among assembled reads, coverage depth as defined here is not the same as the total number of reads aligned in the multiple sequence assembly but as the number of reads with an aligned base that supports the consensus base call. HpaII fragment sequences that did not assemble into multiple sequence alignments (i.e., singletons) were used directly as consensus sequences as well as the Q values calculated by Roche-454's base-calling software.

**Construction of the Paralog Distinguishing List (PDL).** To facilitate the identification of paralogous regions, the MAGIv4.0 C&G database of B73 reference sequences was searched and aligned against itself using BLAT, as described above. All match pairs (not the alignment) with at least 90% identity and a length of 50 bp or longer were used as input for a custom polymorphism detection program. The custom polymorphism detection program performed a Smith-Waterman (Smith and Waterman, 1981) local alignment between match pairs identified by BLAT to obtain a full representation of the alignment in memory. This allowed alignments to be quickly scanned for single base mismatches and single base insertions/deletions (in/dels). Single base mismatches and single base in/dels were identified in the Smith-Waterman local alignments and context sequences were extracted: the 16 bp 5' and 16 bp 3' flanking the mismatch or in/del. All such putative non-allelic differences were extracted as context sequences from all pairwise matches satisfying the 90% identity minimum and 50 bp minimum. These context sequences form the PDL and represent the putative fixed differences that distinguish paralogs. The PDL was used in further analysis to search for paralogous regions, as described below.



**Polymorphism Detection.** Consensus sequences of B73 and Mo17 HpaII fragments were searched against B73 reference sequences (MAGIv4.0 C&G database) using BLAT. Match pairs (not the alignments) were used as input for the custom polymorphism detection program, as described above. Similarly, the polymorphism detection program performed a Smith-Waterman local alignment between the HpaII consensus sequence and the MAGIv4.0 C&G reference sequence (i.e., match pairs) identified by BLAT to obtain a full representation of the alignment in memory. For each single base mismatch or in/del identified by the program, context sequences for B73 and Mo17 HpaII fragment sequences were extracted: the 16 bp 5' and 16 bp 3' flanking the mismatch or in/del. Single base mismatches or in/dels within 16 bp of either end of the local alignment were not considered.

**Implementation of the Paralog Distinguishing List (PDL) and SNP Calling.** With the same custom polymorphism detection program, all context sequences for B73 or Mo17 HpaII fragment sequences were searched against the PDL. Any match to the PDL was considered a paralogous alignment and the entire alignment and all potential SNPs within it were discarded. Otherwise, if no PDL matches were found, all in/del contexts were discarded (not called as SNPs) and the remaining single base mismatch contexts were scanned against a list of SNPs already called. If a single duplicate context was identified in an alignment, only that context was discarded, but if two or more duplicates were identified, the entire alignment was discarded, along with all potential SNPs, even if these SNPs were novel. Provided neither the PDL nor the duplicate alignment check resulted in discarding all potential SNPs, the remaining single base mismatches were called SNPs and no further alignments for the current HpaII consensus sequence were considered. Otherwise, if the alignment was discarded, the next strongest BLAT match was considered, continuing until an alignment was accepted, or until the next

strongest BLAT match was less than 95% identity. This preset 5% maximum was not restrictive for identifying allelic variation, as it is well above the average nucleotide diversity of coding regions between any two maize lines ( $\pi=1-1.4\%$ ) (Tenailon *et al.*, 2001; Wright *et al.*, 2005), but still allows the evaluation of haplotypes that are 5% diverged from one another (Henry and Damerval, 1997). Moreover, the 5% maximum allowed us to use a smaller PDL by avoiding paralogous alignments that were more diverged and easily distinguished from previously reported allelic variation levels. Identified B73/Mo17 putative SNPs and the PDL are available for download from Panzea (<http://www.panzea.org>).

**Panzea SNP Comparison.** We extracted 6,094 B73 and 6,200 Mo17 sequences from the Panzea database (Zhao *et al.*, 2006) that were generated by PCR-directed Sanger sequencing of candidate gene loci. Overlapping sequences that were amplified from the same candidate gene locus were assembled using the procedure described above, except that sequences were clustered based on a common Panzea locus ID. For many of the candidate gene loci, there were two independent amplifications and sequencings of B73 and Mo17 for quality control. This resulted in 3,683 (1.57 Mb) and 3,696 (1.57 Mb) assemblies for B73 and Mo17, respectively. We called SNPs from these sequences using the program already described, except allelic B73 and Mo17 consensus sequences were paired on the basis of common Panzea locus ID. The PDL was not used to call SNPs with Panzea sequences, because it was assumed that all Mo17/B73 pairings were allelic on the basis of single locus PCR amplification. Identified Panzea SNPs were mapped to Mo17 454 consensus sequences on the basis of the 16 bp 5' and 16 bp 3' context sequences, and vice versa, to identify which SNPs from each dataset were called from sequence in common to both datasets. We separately looked at the intersection of Panzea SNPs and B73/Mo17 HpaII SNPs called with (126,683 SNPs; no thresholds) and

without (174,476 SNPs; no thresholds) the PDL. We then compared SNPs that mapped to both datasets to estimate the rate of false SNP discovery and power, assuming that all true Mo17/B73 SNPs were discovered in the Panzea dataset and no false SNPs were discovered.

## 2.3 Results

**Construction of Modified HMPR Libraries.** We modified the previously described HMPR library construction method (Emberton *et al.*, 2005) to allow high-throughput gene-enrichment sequencing of the maize genome using the 454 Genome Sequencer FLX (GS FLX) pyrosequencing instrument (see Materials and Methods). HpaII, a MCS 4 bp cutter (5'-C/CGG-3'), was selected to construct modified HMPR libraries, because of its strong bias for cleaving within unmethylated genic and low-copy regions of the maize genome (Antequera and Bird, 1988; Emberton *et al.*, 2005; Yuan *et al.*, 2002). The first of the two major modifications to the HMPR method was to allow maize genomic DNA to be completely digested with HpaII rather than partially digested. This was done to produce a more repeatable HpaII restriction pattern and, as a result, consistently enrich for gene fragments mostly smaller than 600 bp. Second, HpaII fragments between the sizes of 100-600 bp were gel-isolated and converted via random ligation into concatemers of longer lengths more suitable for nebulization (i.e., fragmentation). At the time of this experiment, it was not possible for us to execute paired-end read sequencing and to routinely obtain read lengths longer than 250 bases on the 454 GS FLX instrument; thus, we used ligation and nebulization in combination to construct and randomly break HpaII concatemers in order to completely sequence larger HpaII fragments.

To test and optimize our library construction method, we constructed modified HMPR libraries for maize inbred lines B73 (husk and root) and Mo17 (root). One concern with modified HMPR and its predecessor is the potential enrichment of organellar genome fragments in constructed libraries (Emberton *et al.*, 2005), as these genomes are unmethylated (Palmer *et al.*, 2003) and, depending on the tissue type, may be present at a very high copy number (Li *et al.*, 2006). Thus, we evaluated as sources of genomic DNA two etiolated tissue types that were expected to have a relatively low abundance of chloroplasts: inner husk leaves (pale green) and dark-grown seedling roots (white). For inner husk leaves, purification of nuclei prior to genomic DNA extraction was used to further limit the amount of co-isolated chloroplast DNA. For dark-grown seedling roots, we used a higher yielding and less laborious total genomic DNA extraction procedure that lacked a nuclei purification step, because dark-grown seedling roots were expected to be highly deficient in chloroplasts and other types of plastids (Possingham, 1980).

**Compositional Analysis of Modified HMPR Libraries.** Modified HMPR libraries and an unfiltered (UF) B73 library were sequenced on the 454 GS FLX instrument (see Materials and Methods). Because the modified HMPR libraries were comprised of randomly concatenated HpaII fragments (see previous section), prior to analysis 454 reads pertaining to these libraries were *in silico* digested with HpaII to produce independent, non-chimeric sequences. To examine the sequence composition of modified HMPR and UF libraries, HpaII fragment and UF sequences were searched against several plant nucleotide databases and genome sequences (see Materials and Methods). The distribution of sequence among these categories is shown in Table 2.2. A higher level of organellar contamination was found in root libraries, but this was offset by their lower level of repeats. B73 and Mo17 root libraries were 7- to 8-fold lower in repeats relative to the B73 husk library, and 14- to 16-fold lower in repeats relative to the UF library. The very low

repeat content of root libraries is comparable to that previously reported in maize HMPR libraries (Emberton *et al.*, 2005) and superior to other non-transcriptome-based gene-enrichment sequencing technologies tested on maize (Gore *et al.*, 2007; Palmer *et al.*, 2003; Rabinowicz *et al.*, 1999; Whitelaw *et al.*, 2003; Yuan *et al.*, 2003). Even though the amount of repeat sequences within modified HMPR libraries varied substantially between tissue types (e.g., B73 husk vs. B73 root), additional biological and technical replications are needed to determine if these differences are attributed to tissue-specific differential methylation of genes and repeats.

The desired enrichment for the genic fraction of the maize genome in root libraries was compromised by an abundance of sequences that did not significantly match any of the screened plant nucleotide databases or genome sequences. These unknown contaminant sequences were most prevalent in the B73 root library, comprising 68.8% of the HpaII fragment sequences. We randomly sampled 1,000 of these putative non-maize sequences from each root library and searched them with BLAST (Altschul *et al.*, 1997) against NCBI's non-redundant nucleotide database. On average, 65% of these sampled sequences had no significant similarity (cutoff E-value of  $10^{-5}$ ) to any sequence with another 30% showing different degrees of similarity to bacterial sequences (results not shown). We suspect that bacterial endo- or exo-symbionts of maize roots were living beneath the seed pericarp layer and subsequently proliferated on seedling roots. Neither the seed surface sterilization procedure nor the sterile seedling growth conditions used in this study would have eliminated any type of bacterial symbiont from seedling roots, thus allowing the co-isolation of bacterial genomic DNA and its enrichment in modified HMPR root libraries. Regardless of the source or identity of these sequences, these putatively non-maize sequences as well as the maize repeat and organellar sequences were excluded from further analyses.

Table 2.2: Sequence composition of the modified HMPR libraries and untreated genomic control libraries (UF). Non-maize sequences were those sequences which did not have a match in any of the screened plant nucleotide, organellar, or repeat databases and are suspected of bacterial origin (see text). Repetitive sequences were determined on the basis of strong homology to TIGR Maize Repeat database version 4.

|               | Modified HMPR |      |      |          |       |      |           |       |      |         |       |      |
|---------------|---------------|------|------|----------|-------|------|-----------|-------|------|---------|-------|------|
| Libraries     | B73 Husk      |      |      | B73 Root |       |      | Mo17 Root |       |      | UF      |       |      |
|               | No.           | Mb   | %    | No.      | Mb    | %    | No.       | Mb    | %    | No.     | Mb    | %    |
| 454 reads     | 391,778       | 65.6 | -    | 470,918  | 101.2 | -    | 1,284,692 | 236.7 | -    | 543,385 | 130.9 | -    |
| Total         | 479,565       | 63.6 | 100  | 771,557  | 97.6  | 100  | 1,937,032 | 225.5 | 100  | 543,350 | 130.9 | 100  |
| Chloroplast   | 3,771         | 0.6  | 0.8  | 5,567    | 0.9   | 0.7  | 30,835    | 4.1   | 1.6  | 3,118   | 0.8   | 0.6  |
| Mitochondrial | 1,319         | 0.2  | 0.3  | 20,332   | 3     | 2.6  | 224,593   | 29.7  | 11.6 | 5,493   | 1.4   | 1    |
| Non-maize     | 6,829         | 0.9  | 1.4  | 530,876  | 67.4  | 68.8 | 454,413   | 49.1  | 23.5 | 41,149  | 9.8   | 7.6  |
| Repeats       | 150,786       | 21.7 | 31.4 | 34,378   | 5.2   | 4.5  | 75,225    | 9.3   | 3.9  | 343,072 | 83.8  | 63.1 |
| Non-repeats   | 316,860       | 40.2 | 66.1 | 180,404  | 21.1  | 23.4 | 1,151,966 | 133.3 | 59.5 | 150,518 | 35.1  | 27.7 |

To assess the degree to which modified HMPR libraries were enriched with genic sequences, we searched non-repetitive, maize HpaII sequences against the Maize Assembled Genome Island version 4.0 Contigs and Singletons (MAGIv4.0 C&S) database (<http://magi.plantgenomics.iastate.edu/>). The MAGIv4.0 C&S database is a partial genome assembly of Sanger-based BAC end and shotgun sequences, gene-enriched genome survey sequences as well as whole-genome shotgun sequences from maize inbred line B73 (Kalyanaraman *et al.*, 2007). In addition, the MAGIv4.0 C&S database represents the most comprehensive maize genomic database in advance of the pending draft maize genome sequence. The search results revealed an intermediate to high intersection (52.2-67.0%) between the MAGIv4.0 C&S database and non-repetitive HpaII fragment sequences contained within modified HMPR libraries (Table 3.3). Moreover, alignment to computationally predicted genes from MAGIv4.0 Contig sequences and the Dana-Farber Cancer Institute (DFCI) maize gene index (<http://compbio.dfci.harvard.edu/tgi/>) showed that modified HMPR libraries were 4- to 5-fold enriched for genes relative to the UF library Table 2.3. This level of gene-enrichment in modified HMPR libraries was similar to that obtained with the original HMPR method (Emberton *et al.*, 2005) and other non-EST-based gene-enrichment sequencing technologies tested on maize (Gore *et al.*, 2007; Palmer *et al.*, 2003; Rabinowicz *et al.*, 1999; Whitelaw *et al.*, 2003; Yuan *et al.*, 2003).

### **Sequence Assembly and Construction of a Paralog Distinguishing List (PDL).**

Why is it challenging to identify SNPs in maize using next generation sequencing technologies? Maize is hypothesized to be an ancient tetraploid (Gaut and Doebley, 1997; Swigonova *et al.*, 2004; Wei *et al.*, 2007), but its genome has lost a substantial number of unlinked duplicated genes (Lai *et al.*, 2004). However, nearly one-third of all maize genes still have a paralog (Blanc and Wolfe, 2004), and many of these paralogs

are tandemly arrayed (Messing *et al.*, 2004). It is estimated, based on ESTs, that maize paralogs resulting from an ancient tetraploid event have diverged a minimum of 10% over time (Blanc and Wolfe, 2004), but recent evidence conservatively suggests that nearly identical paralogs ( $\geq 98\%$  identity) are almost 13-fold more frequent in the maize genome than that of *Arabidopsis* (Emrich *et al.*, 2007). With long enough sequencing reads, unique flanking sequence can be found to distinguish recently diverged paralogs. However, it is unlikely that HpaII fragment sequences, with an average length of 120 bases after *in silico* digestion and a higher single-read error rate than that of Sanger sequencing, will contain sufficient and accurate information to distinguish between highly similar paralogs in the maize genome. In addition, if recently duplicated genes have diverged within the range of previously reported maize nucleotide diversity levels ( $\pi=1-5\%$ ) (Henry and Damerval, 1997; Tenaillon *et al.*, 2001; Wright *et al.*, 2005), it will be difficult, if not impossible, to reliably distinguish paralogs based on the best reference match, reciprocal best match, or a conservative maximum allelic diversity threshold. Finally, the MAGIv4.0 C&S reference database used for SNP calling in this study is a partial genome assembly, thus the true allelic copy for an HpaII fragment sequence may not even be present in this reference database.

Table 2.3: Gene Enrichment of HMPR libraries relative to untreated total genomic DNA libraries (UF).

| Databases                       | B73 Husk |      | B73 Root |      | Mo17 Root |      | UF      |      |
|---------------------------------|----------|------|----------|------|-----------|------|---------|------|
|                                 | No.      | %    | No.      | %    | No.       | %    | No.     | %    |
| MAGIv4.0 contigs and singletons | 244,189  | 52.2 | 131,398  | 61.2 | 822,117   | 67   | 124,323 | 25.2 |
| MAGIv4.0 contigs                | 207,576  | 44.4 | 118,367  | 55.1 | 784,094   | 61   | 87,387  | 17.7 |
| MAGIv4.0 contigs genes          | 129,095  | 27.6 | 75,453   | 35.1 | 501,116   | 40.8 | 41,004  | 8.3  |
| DFCI maize gene index           | 75,027   | 16   | 44,454   | 20.7 | 317,016   | 25.8 | 23,124  | 4.7  |
| Total maize nuclear             | 467,646  | 100  | 214,782  | 100  | 1,227,191 | 100  | 493,590 | 100  |

A two-pronged strategy was developed to deal with some of these challenges. First, the redundant and overlapping non-repeat B73 (husk and root: 61.3 Mb) and Mo17



(root: 133.3 Mb) HpaII fragment sequences (Table 2.2) were assembled into multiple sequence alignments and a consensus sequence representing each alignment was derived. Assembly of these sequences resulted in the derivation of 339,730 (42.6 Mb) and 586,237 (70.7 Mb) non-redundant HpaII consensus sequences from B73 and Mo17, respectively (Table 2.4). In addition to providing a longer assembled sequence to help accurately align HpaII fragments to allelic B73 reference sequences contained within the MAGIv4.0 C&S database (i.e., distinguish between highly similar paralogs), the assembly permitted a calculation of the per-base coverage depth, or the frequency with which any consensus base was observed in the raw data. Importantly, this metric can serve as a measure of confidence in the accuracy of consensus bases, as putative SNPs with a high coverage depth are more likely to be valid (Barbazuk *et al.*, 2007). In addition, the assembly of cognate HpaII fragment sequences reduced the computational requirements for the alignment and SNP calling process, as only unique sequences were used.

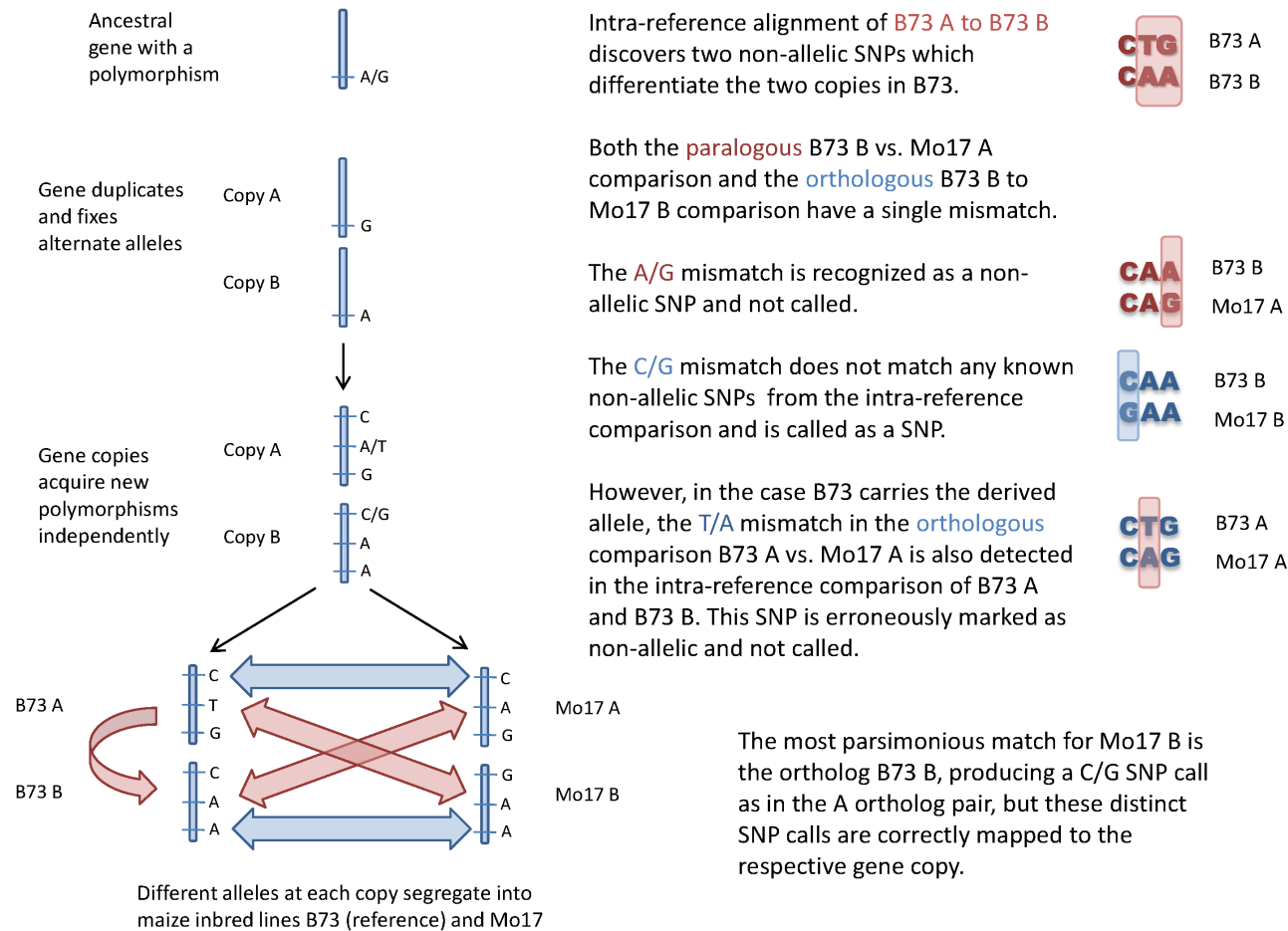
Table 2.4: Summary of the sequence assembly process. Coverage depth refers to the depth of coverage at each individual consensus sequence base, not to the total number of reads aligned in the assembly which may only partially overlap.

| Coverage Depth | B73 Husk and Root |      |      | Mo17 Root |      |      |
|----------------|-------------------|------|------|-----------|------|------|
|                | No.               | Mb   | %    | No.       | Mb   | %    |
| 1              | 263,952           | 31.1 | 77.7 | 415,411   | 42.5 | 70.9 |
| 2              | 44,088            | 6.1  | 13   | 65,846    | 9.1  | 11.2 |
| 3              | 15,188            | 2.3  | 4.5  | 31,473    | 4.8  | 5.4  |
| 4              | 6,745             | 1.1  | 2    | 20,564    | 3.4  | 3.5  |
| 5+             | 9,757             | 2    | 2.9  | 52,943    | 10.9 | 9    |
| Total          | 339,730           | 42.6 | 100  | 586,237   | 70.7 | 100  |

Second, we developed a computational approach to minimize the number of SNPs called from alignments of paralogous sequences, which is similar in objective to the paralog identification method used by the SNP calling software POLYBAYES (Marth *et al.*, 1999) and to the “monoallelism” rules used by Barbazuk *et al.* (2007). Our approach

assumes that it is possible to discover fixed differences among paralogs by comparing a reference sequence database or genome against itself, where almost all sequence differences observed in non-self alignments are non-allelic (Figure 2.1). Although some non-allelic differences may actually be polymorphisms at one or both of the loci, it is assumed that the majority of these identified differences are expected to be fixed differences that distinguish paralogs. Following this argument, a search of the MAGIv4.0 C&S database against itself was performed to identify all such single nucleotide differences that distinguish paralogs in the maize B73 genome. Putative non-allelic fixed differences that were identified from unique paralogous alignments were catalogued into a “paralog distinguishing list” (PDL) in SNP detection power.

**SNP Identification.** With the implementation of the PDL, HpaII consensus sequences from Mo17 were aligned against the best reference match B73 sequence (MAGIv4.0 C&S; 675.2 Mb) and all single nucleotide differences were identified and extracted as context sequences (see Materials and Methods). If the context sequence of any of these single nucleotide differences (Mo17 HpaII vs. B73 MAGIv4.0 C&S) matched a context sequence contained within the PDL, it was treated as an indication of a paralogous alignment and all SNP calls from such alignments were suppressed. In this case, the next strongest alignment for the same HpaII consensus sequence was considered, continuing in this fashion until an alignment with no match to a PDL context sequence was found, or the rate of mismatches in the successive alignments exceeded a preset maximum of 5%. Essentially, the PDL selected which alignments to use for SNP calling but not which single nucleotide differences to call as SNPs. The same procedure was performed with B73 HpaII consensus sequences, which served as an internal control to estimate the rate of false SNP discovery with and without implementation of the PDL.



**Figure 2.1:** Schematic illustrating the paralog detection list methodology in which a comparison of the reference database against itself is used to identify single nucleotide substitutions and single base insertions and deletions that distinguish highly similar but distinct regions of the genome. Presence or absence of these paralog distinguishing “SNPs” in any given alignment of a 454 consensus sequence against the reference database is used to select the correct (orthologous) alignment amongst multiple highly similar alignments. As described in the figure, derived alleles in the reference line are mistakenly identified as paralog distinguishing sites, resulting in the loss of detection of these true SNPs.

Use of the PDL proved to be highly effective at preventing false SNP calls because of paralogous alignments. The estimated false discovery rate (FDR) obtained by comparing the SNP call rate for B73 (control, all SNPs considered false) and Mo17 HpaII consensus sequences at various coverage depths and base quality values (Q values) thresholds is shown in Table 2.5. If SNP calls were made using the PDL and not restricted to a specific coverage depth or Q value threshold, 126,683 putative SNPs between Mo17 and B73 (1 SNP/248 bp) were discovered at an estimated 15.1% FDR. If SNP calls were made using only the most parsimonious alignment (i.e., without PDL), 174,476 putative B73/Mo17 SNPs (1 SNP/199 bp) were called at a dramatically increased FDR of 46.8%. Overall, use of the PDL effectively provided a 3-fold reduction in the rate of false SNP discovery at every evaluated coverage depth and Q value threshold relative to rates determined without use of the PDL.

As shown in Table 2.5, we observed a polymorphism rate of 1 SNP every 216 bp (86,830 SNPs/18,794,000 bp) at an estimated 11% FDR (Coverage Depth:  $\geq 1X$ ; Q-score:  $\geq 35$ ). If we restricted SNP calling to a coverage depth of  $\geq 2X$  (Q-score: all), then we observed a polymorphism rate of 1 SNP every 204 bp at a false SNP discovery rate of 8.4%. The SNP discovery rate for Mo17 HpaII consensus sequences at only 1X coverage (i.e., singletons) and all Q-scores was 1 SNP every 290 bp (calculated from Table 2.5) at an estimated 19.7% FDR, which suggests that at higher coverage depths and with higher quality sequence data more SNPs/kb were captured (i.e., higher SNP detection power). Although the FDR was reduced nearly 2-fold (15.1 to 8.4%) when using the PDL and additionally restricting SNP calls to a coverage depth of  $\geq 2X$ , the FDR remained relatively unchanged at progressively higher coverage depth thresholds. This suggests that deeper sequencing would provide limited improvement in the calling accuracy of SNPs already at a coverage depth of 2X or higher, but this might not have

been the case if the sequenced maize lines were highly heterozygous. The ability to reduce the number of false positive SNPs by restricting SNP calls to higher cover depths was also a key finding by Barbazuk et al. (2007), the first study that used pyrosequencing to identify SNPs within expressed maize genes. Additionally, it seems that Q values calculated by the 454 base calling software (single reads) or CAP3 program (multiple sequence alignments) are of minimal value for eliminating false positive SNPs that result from sequencing errors when SNP calls are restricted to a coverage depth of 2X or higher.

**SNP Validation.** To independently cross-validate a subset of B73/Mo17 HpaII SNPs that were identified via 454 pyrosequencing, we extracted a collection of B73 and Mo17 amplicon sequences from the Panzea database (<http://www.panzea.org/>) (Zhao *et al.*, 2006) that were generated with traditional Sanger sequencing chemistry. The extracted sequences were assembled and aligned according to unique Panzea locus identifiers, which permitted the identification of SNPs. It was assumed that all paired sequences were allelic and all true SNPs were identified (i.e., 0% FDR; 100% power). To estimate an FDR for HpaII SNPs, Panzea SNPs were mapped onto Mo17 HpaII consensus sequences, and vice versa. The mapping resulted in the identification of a subset of SNPs in each dataset that was derived from sequence common to both datasets (Table 2.6).

With the constructed SNP validation dataset, we found that 85.9% (449/523) of the PDL-based HpaII SNPs were concordant with Panzea SNPs. This resulted in an estimated FDR of 14.1%, which strongly agreed with the 15.1% (no thresholds; with PDL) that was estimated using the B73/Mo17 call rate comparison (Table 2.5). However, only 62.0% of SNPs identified in Panzea were also identified in the dataset of PDL identi-

Table 2.5: Summary of putative SNP calls at various base quality and coverage depth thresholds, with and without use of the PDL method described in the text

| CD  | Q   | With PDL |      |         |      |       | Without PDL |      |         |      |        |
|-----|-----|----------|------|---------|------|-------|-------------|------|---------|------|--------|
|     |     | B73      |      | Mo17    |      | FDR   | B73         |      | Mo17    |      | FDR    |
|     |     | SNPs     | Rate | SNPs    | Rate |       | SNPs        | Rate | SNPs    | Rate |        |
| ≥1X | All | 11,904   | 0.61 | 126,683 | 4.03 | 15.1% | 50,936      | 2.35 | 174,476 | 5.02 | 46.8%  |
|     | ≥20 | 10,701   | 0.58 | 119,294 | 4.02 | 14.4% | 47,343      | 2.31 | 164,904 | 5.04 | 45.8%  |
|     | ≥30 | 8,955    | 0.55 | 106,475 | 4.12 | 13.3% | 39,910      | 2.23 | 147,335 | 5.16 | 43.2%  |
|     | ≥35 | 5,703    | 0.51 | 86,830  | 4.62 | 11.0% | 23,149      | 1.92 | 119,465 | 5.74 | 33.4%  |
|     | ≥40 | 2,352    | 0.43 | 62,966  | 4.83 | 8.9%  | 10,378      | 1.78 | 85,547  | 5.92 | 30.1%  |
|     | ≥50 | 1,609    | 0.37 | 57,205  | 4.93 | 7.5%  | 6,832       | 1.46 | 77,688  | 6.03 | 24.2%  |
|     | ≥60 | 879      | 0.32 | 45,610  | 4.88 | 6.6%  | 3,724       | 1.26 | 61,991  | 5.97 | 21.1%  |
|     | ≥70 | 634      | 0.30 | 39,787  | 4.88 | 6.1%  | 2,651       | 1.17 | 54,279  | 5.99 | 19.5%  |
| ≥2X | All | 2,072    | 0.41 | 61,584  | 4.91 | 8.4%  | 9,048       | 1.66 | 83,547  | 6.00 | 27.7%  |
|     | ≥20 | 2,057    | 0.41 | 61,527  | 4.91 | 8.4%  | 9,017       | 1.65 | 83,475  | 6.00 | 27.5%  |
|     | ≥30 | 2,031    | 0.40 | 61,300  | 4.91 | 8.1%  | 8,910       | 1.64 | 83,173  | 6.00 | 27.3%  |
|     | ≥40 | 1,953    | 0.40 | 60,573  | 4.91 | 8.1%  | 8,529       | 1.61 | 82,169  | 6.00 | 26.8%  |
|     | ≥50 | 1,609    | 0.37 | 57,205  | 4.93 | 7.5%  | 6,832       | 1.46 | 77,688  | 6.03 | 24.2%  |
|     | ≥60 | 879      | 0.32 | 45,610  | 4.88 | 6.6%  | 3,724       | 1.26 | 61,991  | 5.97 | 21.1%  |
|     | ≥70 | 634      | 0.30 | 39,787  | 4.88 | 6.1%  | 2,651       | 1.17 | 54,279  | 5.99 | 19.50% |
|     | ≥70 | 634      | 0.30 | 39,787  | 4.88 | 6.1%  | 2,651       | 1.17 | 54,279  | 5.99 | 19.50% |
| ≥3X | All | 702      | 0.33 | 37,980  | 4.88 | 6.8%  | 3,127       | 1.37 | 51,769  | 5.98 | 22.9%  |
|     | ≥20 | 699      | 0.33 | 37,975  | 4.88 | 6.8%  | 3,124       | 1.37 | 51,763  | 5.98 | 22.9%  |
|     | ≥30 | 697      | 0.33 | 37,966  | 4.88 | 6.8%  | 3,114       | 1.37 | 51,751  | 5.98 | 22.9%  |
|     | ≥40 | 689      | 0.32 | 37,912  | 4.88 | 6.6%  | 3,088       | 1.36 | 51,681  | 5.98 | 22.7%  |
|     | ≥50 | 679      | 0.32 | 37,833  | 4.88 | 6.6%  | 3,047       | 1.35 | 51,572  | 5.98 | 22.6%  |
|     | ≥60 | 649      | 0.32 | 37,448  | 4.87 | 6.6%  | 2,899       | 1.32 | 51,044  | 5.97 | 22.1%  |
|     | ≥70 | 529      | 0.30 | 35,417  | 4.87 | 6.2%  | 2,299       | 1.21 | 48,339  | 5.97 | 20.3%  |
|     | ≥70 | 529      | 0.30 | 35,417  | 4.87 | 6.2%  | 2,299       | 1.21 | 48,339  | 5.97 | 20.3%  |
| ≥4X | All | 322      | 0.31 | 24,454  | 4.81 | 6.4%  | 1,452       | 1.31 | 33,403  | 5.90 | 22.2%  |
|     | ≥20 | 319      | 0.31 | 24,454  | 4.81 | 6.4%  | 1,449       | 1.30 | 33,402  | 5.90 | 22.0%  |
|     | ≥30 | 318      | 0.31 | 24,454  | 4.81 | 6.4%  | 1,445       | 1.30 | 33,402  | 5.90 | 22.0%  |
|     | ≥40 | 317      | 0.30 | 24,451  | 4.81 | 6.2%  | 1,443       | 1.30 | 33,399  | 5.90 | 22.0%  |
|     | ≥50 | 316      | 0.30 | 24,443  | 4.81 | 6.2%  | 1,437       | 1.30 | 33,391  | 5.90 | 22.0%  |
|     | ≥60 | 313      | 0.30 | 24,430  | 4.81 | 6.2%  | 1,426       | 1.29 | 33,368  | 5.90 | 21.9%  |
|     | ≥70 | 311      | 0.30 | 24,356  | 4.81 | 6.2%  | 1,405       | 1.28 | 33,272  | 5.90 | 21.7%  |
|     | ≥70 | 311      | 0.30 | 24,356  | 4.81 | 6.2%  | 1,405       | 1.28 | 33,272  | 5.90 | 21.7%  |

fied B73/Mo17 HpaII SNPs, whereas it was 80.9% without the PDL. This signifies a weakness of the MAGIv4.0 C&S-based PDL, as true SNPs were incorrectly considered non-allelic by the PDL.

Table 2.6: Summary of the comparison of SNP obtained from PCR-directed single-locus Sanger sequencing in B73 and Mo17, as extracted from the Panzea database, to SNP calls presented in this study (HpaII) utilizing the PDL method but with no base quality or coverage depth thresholds.

|             | With PDL | Without PDL |
|-------------|----------|-------------|
| Panzea SNPs | 724      | 724         |
| HpaII SNPs  | 523      | 720         |
| Shared SNPs | 449      | 586         |
| HpaII FDR   | 14.1%    | 18.6%       |
| HpaII Power | 62.0%    | 80.9%       |

## 2.4 Discussion

Next generation DNA sequencing technologies have made high-throughput resequencing efficient and affordable. However, the use of these technologies in a read-to-reference based SNP discovery approach at the level of a whole-genome has not come to fruition for agronomically important plant species. The primary reason is that many of these plant species have large, complex genomes and as a result do not have an available, accurate or complete genome sequence. In addition, the short read lengths produced by these high-throughput sequencing technologies are limited in ability to differentiate the large numbers of paralogs that are common to the genome of many angiosperm species (Blanc and Wolfe, 2004). Maize was chosen as the test organism for this pilot study because of three qualities of its nuclear genome: it is  $\approx 2500$  Mb in size; it consists of more than 75% highly repetitive DNA (Meyers *et al.*, 2001; SanMiguel *et al.*, 1996); and at least one-third of its estimated 59,000 genes are duplicated (Blanc and Wolfe, 2004; Messing *et al.*, 2004). Here, we tested a gene-enrichment sequencing approach that is applicable to virtually any plant species and a computational pipeline that enables the efficient and accurate discovery of a large number of SNPs using an incomplete and low-coverage reference sequence.

We modified the previously described HMPR technique (Emberton *et al.*, 2005) to enable shotgun sequencing of 100-600 bp HpaII fragments in a manner that fully used the read length (potential of 200-300 bases) ability of the 454 GS FLX instrument. Of the two tissue types that were tested as sources of genomic DNA, seedling roots have a greater potential to enable the rapid construction of gene-enriched, modified HMPR libraries that have low levels of repeats and organellar DNA contamination. However, improved seed sterilization procedures and/or sterile, antibiotic-treated growing conditions are necessary to prevent the proliferation of bacterial symbionts in seedling roots, and the cytosine methylation pattern of genes and repeats in seedling root tissue needs to be more fully investigated. Since performing this experiment, we have identified unfertilized, immature ear shoots as an excellent tissue for isolating total maize genomic DNA. B73 and Mo17 immature ear HpaII libraries constructed with modified HMPR technology were highly enriched (4-5-fold) for genic sequences, while extremely depleted in repeat, organellar, and bacterial sequences (total: <10%) (M. Gore, R. Elshire, and E. Buckler, unpublished data).

Although our modified HMPR technique facilitated high throughput gene-enrichment sequencing of a large, complex plant genome, in general, the yield per run of modified HMPR libraries on the 454 GS FLX was lower than the expected 100 Mb. If the DNA copy per bead ratio is carefully optimized for modified HMPR libraries, it should be possible to routinely obtain 100 Mb of sequence data. In addition, the low sequencing yield may be because of less than optimal lengths (3-10 kb) of HpaII concatemers. If so, a 6 bp MCS restriction enzyme may help to produce much larger concatemers that are better suited for the downstream 454 sample preparation, which is optimized for undigested total genomic DNA. Also, assembly of the larger restriction fragment sizes would produce larger consensus sequences for more accurate mapping.



Alternatively, with the increased average read length (400 bases) and paired-end read capability of the new GS FLX Titanium (<http://www.454.com>), it might be more efficient and as comprehensive to directly sequence restriction fragments instead of concatemers.

We identified 126,683 putative B73/Mo17 SNPs, primarily in genic regions of the maize genome, using a computational pipeline for short read lengths that is applicable to any plant species with at least a large collection of genome survey sequences. A computational approach was developed to distinguish between allelic and paralogous HpaII consensus-MAGIv4.0 C&S reference alignments by searching identified putative single nucleotide differences against a Paralog Distinguishing List of putative fixed differences that distinguish paralogs from each other. The false SNP discovery rate with implementation of the PDL was estimated by two different approaches, and both were found to be at an acceptable level and highly concordant (15.1 vs. 14.1%). Detection of SNPs using the PDL was 3-fold more effective in controlling the FDR than a most parsimonious alignment strategy, and the FDR could be further reduced by filtering SNPs based on coverage depth and/or Q value thresholds (Table 2.5). The most likely sources of false positive SNPs are cloning artifacts (i.e., base substitution errors) contained within MAGIv4.0 C&S sequences and paralogous alignments not identified by the PDL. Although very stringent parameters were used to assemble redundant, overlapping HpaII fragment sequences, it is possible that collapsed paralogs also contributed to the identification of false positive SNPs. The number of false positive SNPs that result from the FLX system are expected to be low (presumably less frequent at coverage depths of 2X and higher), as other studies have shown the GS FLX single-read error rate to be  $\approx 0.5\%$  (Droege and Hill, 2008) and substantially lower at higher coverage depths (Lynch *et al.*, 2008; Smith *et al.*, 2008). In addition, the rate of paralog collapse in the MAGI assemblies was estimated to be  $\approx 1\%$  (Emrich *et al.*, 2007); therefore, their contribution to the calling of

false positive SNPs and inaccuracies in the PDL should be very minimal.

The difference in FDR estimates between SNPs called with and without the PDL method is much less striking for the Panzea validation dataset (Table 2.6) than that observed for the B73/Mo17 call rate comparison (Table 2.5). This is most likely because Panzea sequences resulted from the preferential sequencing of putatively single-locus PCR products, as PCR reactions that appeared to amplify multiple loci were discarded prior to sequencing (E. Buckler, unpublished). Essentially, the amplicon-Sanger sequencing strategy acted as a PDL. Thus, the Panzea dataset is poorly suited to assess the ability of the PDL to detect paralogous alignments, because the Panzea database was constructed with a bias against paralogous sequences. All amplicon-Sanger sequencing strategies will have this same bias; therefore, the best external validation of the PDL is to sequence modified HMPR libraries of Mo17 on a different next-generation sequencing platform (e.g., Illumina sequencing). Currently, the B73 (internal control)/Mo17 call rate comparison is the best available method to estimate the ability of the PDL to reduce the number of false positive SNP calls from paralogous alignments (Table 2.5). Nevertheless, minor improvements in the FDR are still observed when the PDL is used on the Panzea dataset (Table 2.6).

Transcriptome sequencing is useful when the aim is enrichment of tissue and developmental-stage specific genes; however, for high coverage of the gene space it is not very cost effective. Essentially, numerous cDNA libraries capturing multiple developmental stages and environmental stresses are needed to even approach high coverage of the gene space. Therefore, we sequenced modified HMPR genomic libraries because it is expected to result in a more comprehensive sampling of genes than that

of transcriptome sequencing (Emberton *et al.*, 2005; Palmer *et al.*, 2003), and it is also expected to provide access to the nucleotide diversity in introns, regulatory regions, and non-expressed genes. We used the Lander-Waterman model (Lander and Waterman, 1988) and the rate of contig formation as described in Whitelaw *et al.* (2003) to estimate the effective gene space size sampled by the modified HMPR method, which was 136.4 Mb ( $\approx 27\%$  of the  $\approx 500$  Mb maize gene space; Palmer *et al.*, 2003) for the Mo17 root library. This estimate of the effective gene space size might be slightly overestimated due to the very stringent CAP3 assembly parameters that were used. Given that 70.7 Mb of HpaII consensus sequence data exists for Mo17 (Table 2.4), it is estimated that the library was sequenced to only 0.52X coverage. If we were to sequence the Mo17 root library to 1X coverage, then the maximum number of putative SNPs called with the PDL would be  $\approx 200,000$  at a rate of 4.03 SNPs/kb. If several million SNPs are to be discovered, we will need to sequence additional maize inbred lines, possibly construct other modified HMPR libraries using different 4 bp cutter MCS restriction enzymes, and/or use the draft maize genome sequence to call SNPs.

The PDL is only as high-quality as the completeness and accuracy of the reference sequence used to construct it, but despite the shortcomings of the MAGI assemblies (e.g., 1% collapsed paralogs, cloning artifacts, and partial genome assembly), a significant reduction (3-fold) in the number of false positive SNPs that resulted from paralogous alignments was still observed (Table 2.5). Moreover, these issues will be mostly resolved when the draft maize B73 genome sequence is available for constructing a PDL and calling SNPs.

A more important limitation of the PDL, however, is that it reduced the power to

detect true SNPs. Based on the observed SNP call rate (4.91 SNPs/kb; 1 SNP/204 bp) with the PDL at a coverage depth of  $\geq 2X$ , we are under-estimating the expected SNP call rate (1 SNP/153 bp based on 1,095 genes) between any randomly chosen diverse, temperate maize inbred lines by  $\approx 25\%$  (Yamasaki *et al.*, 2005). If SNPs were called without the PDL at a coverage depth of  $\geq 2X$ , the observed (6.00 SNPs/kb; 1 SNP/167 bp) and expected (1 SNP/153 bp) SNP call rates are nearly identical. As shown in Table 2.6, based on the comparison of B73/Mo17 HpaII SNPs (no threshold) with the Panzea SNP dataset, there was an 18.9% loss in SNP detection power with implementation of the PDL. The reduction in power is attributed to true SNPs being incorrectly considered non-allelic by the PDL. We hypothesize that these true SNPs could not be distinguished from actual fixed differences among paralogs on the basis of the intra-reference sequence comparison alone, which would occur if the reference line (B73) used to construct the PDL carries a derived allele (Figure 2.1). This is a systematic bias that may affect both population genetic and association studies when the reference line alone carries an allele of interest. This problem is most severe when a single line is compared to the reference, but the expected rate of false negatives because of this effect decreases to  $1/(n+1)$  when  $n$  lines are compared to the reference. Further reduction may be possible if multiple non-reference lines are also compared to each other.

Although the results obtained in this pilot study are very encouraging, there are several drawbacks to this approach that should be considered. First, the method of gene enrichment used here restricts SNP discovery to sites near HpaII restriction sites in unmethylated regions, which can be remedied by constructing additional modified HMPR libraries with different 4 bp cutter MCS restriction enzymes. We do not presume that all nucleotide variation in methylated regions of the maize genome is phenotypically irrelevant, so different methods are needed to discover SNPs from these regions. Additionally,

genome wide methylation patterns and locus specific methylation levels may vary across genetic backgrounds, tissue types, developmental stages, and even environmental conditions (Cervera *et al.*, 2002; Finnegan *et al.*, 2000; Lister *et al.*, 2008; Rabinowicz *et al.*, 1999; Vaughn *et al.*, 2007). Thus, performing this technique across a panel of inbred lines may not result in representation of all lines at all loci. For marker discovery, this line-specific or locus-specific censoring effect may not be important overall, but population genetic studies may be adversely affected by non-random missing data.

Regardless of these limitations, a considerable number of SNPs were discovered at an acceptably low FDR for the purpose of constructing high density multiplexed genotyping products, but sequencing of additional maize inbred lines is needed to construct a SNP dataset with low ascertainment bias that is appropriate for phylogenetics or population genetics studies. However, the SNPs identified in this study are immediately applicable for fine mapping of complex traits in the Intermated B73 x Mo17 (IBM) population, which is a widely used community resource for QTL mapping studies in maize (Lee *et al.*, 2002). Most importantly, we estimate the cost of SNP discovery in this study at \$0.38/SNP yet note that several aspects of the molecular methods used here can be optimized for much higher sequencing yield and broader genome coverage. Such optimization, combined with further advances in high throughput sequencing yield, longer read lengths, lower error rates, and cheaper run costs, can further reduce the cost of SNP discovery in diverse maize such that several million gene-enriched SNPs needed for comprehensive association studies is an immediate economic possibility.

## CHAPTER 3

# ALCHEMY: A RELIABLE METHOD FOR AUTOMATED SNP GENOTYPE CALLING FOR SMALL BATCH SIZES AND HIGHLY HOMOZYGOUS POPULATIONS

Accepted pending minor revision for publication in *Bioinformatics*. M.H. Wright, C.W. Tung, K. Zhao, A. Reynolds, S.R. McCouch, and C.D. Bustamante. *ALCHEMY: A Reliable Method for Automated SNP Genotype Calling for Small Batch Sizes and Highly Homozygous Populations*. *Bioinformatics*, accepted 2010.

**Abstract.** The development of new high-throughput genotyping products requires a significant investment in testing and training samples to evaluate and optimize the product before it can be used reliably on new samples. One reason for this is current methods for automated calling of genotypes are based on clustering approaches which require a large number of samples to be analyzed simultaneously, or an extensive training data set to seed clusters. In systems where inbred samples are of primary interest, current clustering approaches perform poorly due to the inability to clearly identify a heterozygote cluster. As part of the development of two custom SNP genotyping products for *Oryza sativa* (domestic rice), we have developed a new genotype calling algorithm called “ALCHEMY” based on statistical modeling of the raw intensity data rather than modelless clustering. A novel feature of the model is the ability to estimate and incorporate inbreeding information on a per sample basis allowing accurate genotyping of both inbred and heterozygous samples even when analyzed simultaneously. Since clustering is not used explicitly, ALCHEMY performs well on small sample sizes with accuracy exceeding 99% with as few as 6 samples.

### 3.1 Introduction

The number of single nucleotide polymorphisms which can be genotyped in a single experiment has increased exponentially in the past 5 years, with costs per data point declining at the same time (Maresso and Broeckel, 2008; Kim and Misra, 2007). This technological advance has been critical to the design and execution of cost-effective genome wide association studies (GWAS) in humans and other well studied systems (Hirschhorn and Daly, 2005; McCarthy *et al.*, 2008). While most “catalog products” offered by companies such as Illumina and Affymetrix are developed for human genotyping, the underlying technologies of the assays themselves and the manufacturing methods which produce such high density products should be transferable to most diploid systems of interest and are currently being adapted for domesticated plants and animals.

Development of a custom genotyping product is still an expensive process, especially if re-sequencing for SNP discovery must be performed. Even with a sufficient SNP database on hand, the development of a working assay may require dozens or even hundreds of samples to be run in order to identify which array features are working reliably and which simply do not perform well in the multiplexed environment. Human genotyping products from Affymetrix and Illumina, now in their 5th generation or later, are largely free of SNPs and probes which did not “convert” to working assays, as previous generation products have identified these problem SNPs empirically and they have been removed from later generation products. However, a first generation custom product may see up to 50% or more of the intended SNP assays fail to generate accurate results, and this may only be determined after 100 or more samples have been run. Depending on the number of samples planned for the entire experiment, the cost of the samples needed for development and quality-control procedures for custom genotyping

products may easily form a significant fraction of the total experiment cost.

One limitation in custom genotyping array development is the requirement of many automated genotype calling algorithms such as Affymetrix's "BRLMM-P" to have a large number of samples from which three distinct clusters of genotypes (AA, AB, BB) can be reliably identified and clearly distinguished (Affymetrix Inc., 2006; Rabbee and Speed, 2006; Teo *et al.*, 2007; Carvalho *et al.*, 2007) . The methodology in many of these clustering algorithms implicitly assumes the existence of all three clusters. Other published methods attempt to statistically test whether or not two or three clusters best describes the data (Liu *et al.*, 2003). Some more recent methods such as "Birdseed" (Korn *et al.*, 2008) require 100 or more samples with known genotypes to be assayed in advance to "train" the algorithm. The BRLMM-P algorithm can accept training samples as "priors" or can be run without priors for *de novo* calling. For well-funded studies such as the human HapMap (The International HapMap Consortium, 2005, 2007), it is possible to obtain this prior information and then apply it to future samples. In smaller projects however, obtaining verification data for this many samples may be prohibitive.

Another aspect of high-throughput SNP genotyping assay design is the laboratory protocols used to prepare samples prior to actual genotyping step performed by the manufacturer's system. The Affymetrix human genotyping products have long employed a genome-reduction step where the genome is enzymatically digested and the digestion products ligated with universal adapters followed by PCR amplification of small (<2 kb) fragments. In principle, this method is generalizable to other genomes but may require optimization of the restriction enzymes used and fragment sizes amplified. An alternative, especially for systems with smaller and less complex genomes, is to skip the



complexity reduction step and directly label a randomly digested genome, or amplify with random labeled primers. These options, and others, can only truly be assessed by running some samples and assessing the genotype call rates and accuracies. However, if the calling method is inaccurate with less than 100 samples or requires priors from known genotype samples, experimenting with and optimizing the sample preparation and labeling step is simply too costly. For new and custom products, it is desirable to have a genotype calling method which does not require prior information or training samples and can produce accurate results with only a few samples.

Another consideration not addressed by genotype calling algorithms designed for human applications is the possibility that the samples genotyped may be inbred or deficient in heterozygote genotypes. Many animal model systems have developed panels of inbred lines or strains that are widely used in genetic experiments (Yang *et al.*, 2009). Likewise, in plant systems, many research systems and many agronomically important species have large collections of inbred lines which form the basis of breeding programs and large quantitative genetic studies (Yu *et al.*, 2008; Buckler *et al.*, 2009). Often, these large collections of inbred lines are genetically and phenotypically diverse and thought to capture a large proportion of the naturally occurring variation in these species.

Genotyping these inbred panels presents a possible problem for automated genotype calling based on cluster analysis because the heterozygote cluster, which is always expected for a population in Hardy-Weinberg equilibrium (HWE) if the SNP is segregating, may have very few observations or be completely absent, for nearly all SNPs. In our experience, these deviations from HWE results in very few or no heterozygote samples within the batch and this confuses current software causing one of the homozy-

gous clusters to be declared heterozygous, or one homozygous cluster to be split into heterozygote and homozygote calls. This is related to the problem of requiring large batch sizes since the main problem with analyzing only a few samples at a time is that one or two genotype classes may be completely absent or have too few observations for clustering analysis to reliably identify cluster locations and boundaries.

In the development of two SNP genotyping products for cultivated rice (*Oryza sativa*), we encountered both problems, particularly that posed by the lack of heterozygosity in our largely inbred sample collection. To address this, we developed a custom genotype calling algorithm called ALCHEMY, specifically designed to perform *de novo* calling without prior information and to perform reliably on small numbers of samples while still gaining in accuracy and call rates when multiple samples are available for simultaneous analysis.

## 3.2 Approach

The central idea behind the ALCHEMY algorithm is that the summary raw intensities for each channel (allele) is a mixture distribution composed of a signal component and a noise component. When an allele is present, an intensity value drawn from the “signal” component is observed. If the allele is not present, the intensity value observed is drawn from the “noise” component. Conceptually, under this model, a diploid organism with an AA genotype would have a signal observation on the A channel and a noise observation on the B channel, and likewise but reversed for a BB genotype. The heterozygous genotype AB would produce signal on both channels. The opposite, noise observed on

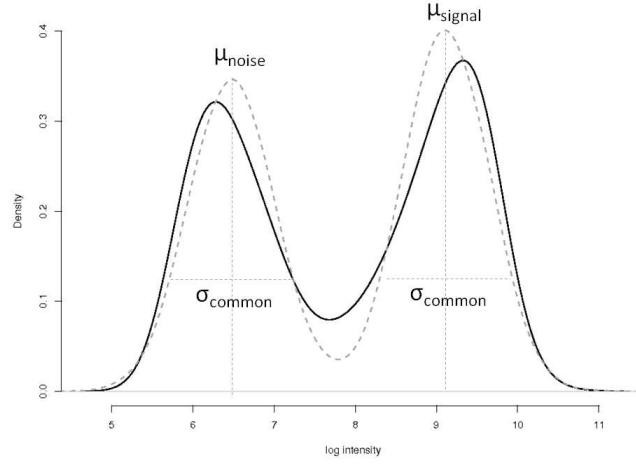


Figure 3.1: Density plot of log intensities across all A allele probes for one sample (black solid line) and fit of Gaussian mixture distribution (gray dashed line)

both channels, indicates an assay failure (no call). This may occur for many reasons, but the two most likely reasons are complete lack of the genomic region (deletion) in the sample, or polymorphism within the flanking sequence which causes non-allele-specific interference with primer or probe binding.

As shown in figure 3.1, these signal and noise modes are readily identified visually and reasonably well approximated by the fit of a Gaussian mixture distribution. Let  $\mu_{s_A}$  and  $\mu_{n_A}$  be the means of the signal and noise distributions for the A channel respectively, and  $\sigma_A^2$  be their common variance. Let  $\pi(AA)$  be the prior probability of observing an AA genotype. Then, using Bayes rule, the posterior probability of AA, AB, BB, or no call (NC) given the observed intensities  $x_A$  and  $x_B$ , is

$$\begin{aligned}
 P(AA|x_A, x_B) &= \frac{P(x_A|\mu_{s_A}, \sigma_A^2)P(x_B|\mu_{n_B}, \sigma_B^2)\pi(AA)}{P(D)} \\
 P(AB|x_A, x_B) &= \frac{P(x_A|\mu_{s_A}, \sigma_A^2)P(x_B|\mu_{s_B}, \sigma_B^2)\pi(AB)}{P(D)} \\
 P(BB|x_A, x_B) &= \frac{P(x_A|\mu_{n_A}, \sigma_A^2)P(x_B|\mu_{s_B}, \sigma_B^2)\pi(BB)}{P(D)}
 \end{aligned} \tag{3.1}$$

$$P(NC|x_A, x_B) = \frac{P(x_A|\mu_{n_A}, \sigma_A^2)P(x_B|\mu_{n_B}, \sigma_B^2)\pi(NC)}{P(D)}$$

where the likelihood terms of the form  $P(x_A|\mu_{s_A}, \sigma_A^2)$  is given by the Gaussian density function.  $P(D)$  represents the total probability of the data which is calculated by summing the numerators of all 4 cases above. For the prior terms, let  $p$  be the frequency of the A allele in the population,  $f_i$  be the inbreeding coefficient for sample  $i$ , and  $z$  be an *a priori* probability of any particular SNP assay failing. Let  $\pi$  be specified by the Hardy-Weinberg equilibrium genotype frequencies adjusted for inbreeding:

$$\pi(AA) = (p^2(1 - f_i) + pf_i)(1 - z) \quad (3.2)$$

$$\pi(AB) = 2p(1 - p)(1 - f_i)(1 - z)$$

$$\pi(BB) = ((1 - p)^2(1 - f_i) + (1 - p)f_i)(1 - z)$$

$$\pi(NC) = z$$

We take as the genotype call the genotype with highest posterior probability. This framework not only provides a conceptually simple means to call genotypes derived directly from first principles and easily verified properties of the data, it also provides an easily understood quality metric for each call. In many other methods, quality metrics are available but their scale is not well defined or easily understood. The posterior probability produced by the ALCHEMY model may be taken directly as the (subjective) probability that the call is correct. In practice, a threshold for this probability of correctness is set and all calls which are below this threshold are taken as “no calls”. This allows a simple trade-off between completeness of the data set produced, and accuracy.

The signal and noise means, as well as standard deviations, for each channel, are

estimated across samples for each SNP independently. This is performed using the Expectation-Maximization (EM) algorithm. Likewise, the allele frequency at each locus is initially assumed to be 0.5 or an *a priori* specified value for each SNP if the user has prior knowledge of this information, then updated via EM. Worthy of note is that the prior distribution  $\pi$  is different for each sample due to the dependence on  $f_i$ . This parameter is also estimated from the data via EM, as described below.

The above discussion, for purposes of clarity, illustrates the central idea behind ALCHEMY and is an accurate description of the initial implementation of the algorithm. However, a few further observations from the data improve the calls obtained and the accuracy of the posterior probabilities associated with these calls. Namely, it is easily seen that the tails of the two components of the mixture distribution in Figure 3.1 are heavier than a Gaussian distribution. Thus, the distribution is better modeled as a mixture of student's  $t$  distributions. Second, it is readily seen (data not shown) that intensity values on the two channels are correlated, particularly for heterozygotes. Rather than take the product of the two channel likelihoods as given above, which is correct only if the two channels are independent, we model both channels together as a bi-variate  $t$ -distribution with fixed correlation between the intensities on each channel. Thus, the numerators in equation 3.1 (for  $P(AA)$  shown below, other cases omitted) are given by

$$P(AA|x_A, x_B) = \frac{(1 - \rho^2)^{\frac{1}{2}}}{2\pi\sigma_A\sigma_B} \left( 1 + \frac{q_A^2 + q_B^2 - 2\rho q_A q_B}{v(1 - \rho^2)} \right)^{\frac{-(v+2)}{2}} \frac{\pi(AA)}{P(D)} \quad (3.3)$$

where,

$$q_A = \frac{x_A - \mu_{s_A}}{\sigma_A} \quad q_B = \frac{x_B - \mu_{n_B}}{\sigma_B}$$

and  $x_A, x_B, \mu_{s_A}, \mu_{n_B}, \sigma_A, \sigma_B$  are as defined previously and  $\rho$  is the correlation between channels, and  $\nu$  the degrees of freedom of the bi-variate t-distribution. As before, signal and noise means ( $\mu_{s_A}, \mu_{s_B}, \mu_{n_A}, \mu_{n_B}$ ) and the common standard deviations of each channel's signal and noise mixture distribution ( $\sigma_A, \sigma_B$ ) are estimated from the data, but  $\rho$  (for each genotype case) and  $\nu$  are fixed parameters that may be adjusted by the user (see methods). Finally, for the Affymetrix and Illumina technologies considered here, the intensity level is typically proportional to the amount of allele present and thus the signal on either channel is reduced for heterozygotes compared to homozygotes. To account for this, we used a reduced value for  $\mu_{s_A}$  and  $\mu_{s_B}$  in the likelihood for  $P(AB)$ . These reduced values may be either a fixed proportion of the homozygote signal levels or estimated from the data if a sufficient number of heterozygotes are observed at a SNP. For all analyses presented here, the heterozygote signal mean parameters were always determined from homozygote signal levels and not estimated from actual heterozygote observations, regardless of the number of heterozygotes observed (see methods).

### 3.3 Algorithm

**Input.** The input to ALCHEMY is the summary intensity values for the A allele and B allele channels, for all SNPs interrogated by the assay. Prior to running ALCHEMY, the actual raw data which is platform dependent is converted into these summary intensities. For Affymetrix arrays which have multiple probes per allele per SNP, the log intensity values for each allele are averaged across probes to create a single summary value. For

Illumina GoldenGate, intensity values are already summarized as one number for each allele as outputted from BeadStudio. The log of the raw (not normalized) intensity is used.

**Normalization.** All intensity values for each sample are normalized to the average total intensity of all input samples. Specifically, the summary values for each channel are summed for each SNP and the mean and standard deviation of the values across SNPs determined. The mean of the sample means and mean of standard deviations is determined, and then each sample adjusted such that the mean and standard deviation of total intensities across SNPs equals this overall mean. For an individual SNP, the sum intensity is adjusted and then divided back into to A and B channel components in proportions equal to the original values.

**EM starting distribution.** For each sample, all intensity values are used to fit a bimodal Gaussian mixture to identify signal and noise means specific to each sample. After analyzing each sample, the mean of the signal means, the mean of noise means, and the mean of the shared standard deviations is retained to parametrize a distribution from which SNP specific parameters are drawn for initial values in the EM algorithm.

**Expectation-Maximization.** For each SNP, random values for the parameters  $\mu_{s_A}, \mu_{n_A}, \mu_{s_B}, \mu_{n_B}, \sigma_A^2, \sigma_B^2$  are drawn from the distribution determined in the step above. Given these values, the probability of each possible genotype call (AA, AB, BB, and no call) is calculated via equation 3.3. The call with maximum posterior probability is assigned and then the maximum likelihood estimates of the model parameters, assuming these genotype calls, are computed. Using these new parameters, genotype call

probabilities are recomputed followed by re-estimation of the model parameters. This continues in an iterative fashion until the genotype call for each sample remains fixed across successive iterations. The genotype call with maximum posterior probability is the final call produced by the algorithm and the posterior probability itself a quality metric which may be subjectively interpreted as the probability the genotype call is correct. In tandem with the Expectation-Maximization step, the A allele frequency  $p$  used in the prior of the Bayesian model is also re-estimated based on the genotype calls at each iteration. Initially,  $p$  is set to 0.5 or a value optionally specified (per SNP) as input to the program.

**EM estimation of inbreeding coefficient.** An estimate or prior belief for the inbreeding coefficient for each sample may be specified as input to the program and in this case this value will be used when calculating the prior distribution ( $\pi$ ) in equation 3.2 and posterior probabilities via equation 3.3. Alternatively, ALCHEMY can estimate the inbreeding coefficient for each sample via Expectation-Maximization. In this case, a randomly selected user-specified number of SNPs are called via the full ALCHEMY algorithm, initially with random values selected for the inbreeding coefficient. Given the genotype calls produced by ALCHEMY, the heterozygosities for each sample are computed ( $H_{obs}$ ) and compared to the heterozygosity expected ( $H_{exp}$ ) given the allele frequency at each SNP (estimated from the current genotype calls) assuming Hardy-Weinberg equilibrium. A new inbreeding coefficient is estimated by  $F = 1 - H_{obs}/H_{exp}$ . Using the new inbreeding coefficient values, genotype calling is repeated for the subset of SNPs. Iteration stops when the improvement in the total likelihood of the data no longer improves or a preset maximum number of iterations is exceeded. EM is performed for a user-specified number of random starting points and the inbreeding coefficients which produced the maximum total likelihood across all samples is retained and



used for the final, full ALCHEMY run on all SNPs.

### 3.4 Methods

**SNP Arrays.** We designed two multiplexed high-throughput SNP genotyping products for use in genotyping a collection of inbred lines of *Oryza sativa* (domestic rice). The first product is an Illumina 1,536 SNP GoldenGate Oligo Pool Assay (OPA) (Fan *et al.*, 2003) intended for use in breeding applications. The second product is an Affymetrix 44,100 SNP GeneChip (Matsuzaki *et al.*, 2004) designed through the company's custom genotyping program. This higher density array is intended both for direct use in GWAS in rice as well as a pilot array for designing a much higher density Affymetrix GeneChip. The vast majority of SNPs for both products were selected from the *Oryza*SNP project's Perlegen resequencing of 20 diverse *Oryza sativa* inbred lines (McNally *et al.*, 2009), selected to represent four of the five major rice sub-populations plus one line from the aromatic/Group V subpopulation (Garris *et al.*, 2005). For both arrays, SNPs were chosen primarily to obtain uniform density across the entire genome and to maximize informativeness both within and between the 4 major subpopulations for which multiple lines were resequenced in the *Oryza*SNP project. As only one line was resequenced in the Aromatic/Group V subpopulation, SNPs private to this subpopulation were not available for selection. Additionally for the 44,100 SNP array, SNPs were chosen to minimize pairwise linkage disequilibrium.

**Samples.** At the time of writing, both products have been utilized on a much larger number of samples than that which is presented in this paper, as part of an ongoing

effort. For the purposes of illustration of ALCHEMY and for consistency and comparability between the data sets, a subset of 166 samples were selected that were run on both the Illumina and Affymetrix platforms with some samples run multiple times on one or both platforms. Counting replicate assays, a total of 200 Affymetrix 44K assays and 184 Illumina 1,536 assays are used. In the Affymetrix data set, 23 samples were run at least twice. In the Illumina data set, 7 samples were run at least twice. The sample selection includes the two rice reference genome lines “Nipponbare” (*temperate japonica*) and “9311” (*indica*), as well as the Nipponbare x 9311 F1, run in several replicates on both platforms. An additional 10 lines have Illumina Genome Analyzer II short-read re-sequencing data which can be used to verify genotype calls. The remaining samples are all inbred domestic rice varieties representing all five major sub-populations of *Oryza sativa* (Garris *et al.*, 2005) and are representative of a typical sample collection of interest.

**QC filtering.** As with any new genotyping product, a number of intended SNP assays fail to convert to working assays in the multiplexed environment for reasons which cannot always be determined. All results presented here for the two rice arrays exclude up front SNPs which did not convert to working assays, to the best of our ability to determine. For the Illumina 1,536 GoldenGate OPA, 114 SNPs were found to be in regions of multiple copy in the genome and another 63 SNPs generated consistently poor posterior call probabilities such that the expectation of error across all samples at these SNPs was >10%. The remaining 1,359 SNPs were evaluated for accuracy and concordance in this study.

For the Affymetrix 44,100 SNP array, 1,127 SNPs were erroneously designed in

multiple copy regions of the genome, 2,280 SNPs generated consistently poor posterior call probabilities (>10% expected error), and an additional 393 SNPs were consistently discordant with all validation data indicating that the assay was not interrogating the intended target or possibly interrogating multiple targets. 40,300 SNPs were evaluated for accuracy and concordance in this study. For both Illumina and Affymetrix, the SNPs excluded by these criteria were excluded from both ALCHEMY and the vendor's software when computing accuracies, concordances, and call rates.

**Run-time Options.** ALCHEMY has several options which control its performance and execution time. Of note, the number of degrees of freedom for the bi-variate t-distribution was set to 7 and the *a priori* probability of individual assay failure set to 0.01. When an EM search was used to optimize inbreeding coefficients, 10 random starting points were selected and EM conducted for a maximum of 10 iterations. For the Affymetrix array, 2000 SNPs were evaluated in the EM search for inbreeding coefficients. For Illumina, all 1,536 SNPs were used. For human HapMap samples, 2000 SNPs were used for each of the NspI and StyI chips which compose the Affymetrix 500K GeneChip. EM searches for optimal parameters for each SNP were conducted with 50 random starting points and a maximum of 20 iterations per starting point before terminating the EM algorithm. The number of EM starting points is the primary determinant of execution time. ALCHEMY is a multithreaded application capable of utilizing multiple CPU execution cores for parallel processing. Utilizing 8 CPU cores and 50 EM starting points for EM searches, ALCHEMY obtains slightly faster run times than BRLMM-P (not a multi-CPU capable program) on the same data. Reducing the number of EM starting points to 10 obtains very similar results to those presented here, but takes 20% of the run time. In all cases, the heterozygote signal mean parameter was set to  $\mu_n + (\mu_s - \mu_n) / \sqrt{2}$ , where  $\mu_n$  is the estimated noise mean and  $\mu_s$  is the estimated signal

mean of the homozygote. The correlation between intensities for AA homozygotes is fixed by the program to  $\tan^{-1}(\mu_{n_B}/\mu_{s_A})$  and analogously for the BB homozygote. The heterozygote correlation is set to  $\tan^{-1}(\mu_{A_{hs}}/\mu_{B_{hs}})$  if  $\mu_{A_{hs}} < \mu_{B_{hs}}$ , and  $\tan^{-1}(\mu_{B_{hs}}/\mu_{A_{hs}})$  if  $\mu_{B_{hs}} < \mu_{A_{hs}}$ , where  $\mu_{A_{hs}}$  is the heterozygous signal mean of the A allele channel, and likewise for the B channel. The covariance matrix of the bi-variate t-distribution for evaluating the likelihood and computing the posterior probabilities is determined by these correlations and the EM estimates of the marginal variances.

### 3.5 Results

In the development of the two genotyping products, we have four types of samples which can be used to evaluate performance of the assays themselves and the ALCHEMY genotype calling method: (1) reference samples, (2) replicate samples, (3) OryzaSNP samples, and (4) samples which have been re-sequenced by high-throughput short-read sequencing (Illumina GenomeAnalyzer II) to a sufficient extent to determine the allele at a large majority of SNP sites. Except where explicitly stated otherwise, all validations were performed comparing either ALCHEMY or the vendor's software (BeadStudio and BRLMM-P) run on the entire collection of 184 Illumina assays or 200 Affymetrix assays.

The “reference” samples are the two rice lines for which assembled genome sequence is publicly available. The first, “Nipponbare”, is a *temperate japonica* line which has been extensively sequenced and assembled into high-quality pseudomolecules (Goff *et al.*, 2002). The second, “9311”, is an *indica* variety which has been sequenced

by Sanger whole-genome shotgun sequencing and assembled using the Nipponbare genome sequence as a scaffold (Yu *et al.*, 2002). Our samples bear the same name as these reference genomes but are not identical to the lines sequenced. As seen in Table 3.1, ALCHEMY calls replicate the expected calls based on the Nipponbare sequence to a high degree, but diverge from the 9311 genome sequence. However, since we have many replicates of these samples, we find that there is high concordance across our reference samples and it seems likely that the differences seen between our 9311 and the genome sequence reflect true differences resulting from different origins of the materials. Likewise, the F1 genotypes which we predict from the Nipponbare and 9311 genome sequences also show differences with the ALCHEMY calls as a consequence of the divergent 9311 lines. However, Mendelian consistency of the Nipponbare, 9311, F1 trio is 99.9%, suggesting again that differences are not due to inaccurate ALCHEMY calls. Regardless, both BeadStudio and BRLMM-P perform much worse than ALCHEMY in both accuracy and call rate and show lower concordances across replicate samples.

Next we looked at the samples run in replicate, including the reference samples above. In the Illumina samples we have 7 samples run at least twice (including the 3 reference samples) with all pairs of replicates showing an average concordance of 99.5% and average pairwise mutual call rate (genotype called in both samples) of 96.8% (Table 3.2). This is also seen in the Affymetrix samples where 23 different samples run at least twice have an average pairwise concordance of 99.7% and average mutual call rate of 89.8%. This indicates both the assays themselves and ALCHEMY genotype calls are consistent across many distinct samples.

We also looked at the concordance between ALCHEMY calls and the *Oryza*SNP

Table 3.1: Comparison of reference lines to published genome sequence. <sup>1</sup>Numbers reported are averages across replicate samples. <sup>2</sup>Percentage of genotype calls which agree with published sequence presuming homozygosity <sup>3</sup>9311 line genotyped in this study was obtained from a different source than the sequenced line (see text). <sup>4</sup> Genotypes predicted from parental genome sequence assuming normal Mendelian transmission and presuming homozygosity of the parents.

| Illumina 1,536 SNP GoldenGate OPA |                |                        |           |                        |           |
|-----------------------------------|----------------|------------------------|-----------|------------------------|-----------|
| line                              | # <sup>1</sup> | ALCHEMY                |           | BeadStudio             |           |
|                                   |                | agreement <sup>2</sup> | call rate | agreement <sup>2</sup> | call rate |
| Nipponbare                        | 7              | 99.2%                  | 98.7%     | 95.7%                  | 99.3%     |
| 9311 <sup>3</sup>                 | 7              | 95.7%                  | 98.2%     | 92.5%                  | 98.5%     |
| NPx9311 <sup>4</sup> F1           | 6              | 94.2%                  | 96.0%     | 89.8%                  | 99.7%     |
| average                           |                | 96.6%                  | 97.8%     | 93.0%                  | 99.1%     |

| Affymetrix 44K GeneChip |                |                        |           |                        |           |
|-------------------------|----------------|------------------------|-----------|------------------------|-----------|
| line                    | # <sup>1</sup> | ALCHEMY                |           | BRLMM-P                |           |
|                         |                | agreement <sup>2</sup> | call rate | agreement <sup>2</sup> | call rate |
| Nipponbare              | 7              | 99.6%                  | 97.5%     | 90.2%                  | 73.4%     |
| 9311 <sup>3</sup>       | 5              | 96.8%                  | 94.9%     | 82.6%                  | 74.1%     |
| NPx9311 <sup>4</sup> F1 | 6              | 96.4%                  | 91.8%     | 81.6%                  | 81.6%     |
| average                 |                | 97.9%                  | 94.9%     | 85.3%                  | 76.3%     |

project's Perlegen sequence from which these assays were designed. Unfortunately, while the materials utilized in our study are identical or as closely related as possible to the original *Oryza*SNP lines, the Perlegen data set contains many missing observations and an average per-line error rate of approximately 2.9% (McNally *et al.* (2009) - MBML intersect set). Comparing ALCHEMY genotype calls on these samples to the Perlegen sequence confirms this with an average concordance of 97.1%.

Finally, in an effort to discover more SNPs for the production of an even larger genotyping array, we have performed short-read next generation sequencing on 8 inbred rice lines utilizing the same material as that which was genotyped. Additionally, another 2 lines have been re-sequenced by a collaborating group (RiceCAP) utilizing materials derived from the same original sources as our materials. Combining these data sets

Table 3.2: Pairwise concordance for replicate samples. <sup>1</sup>Call rate in this table refers to the percentage of SNPs called in both samples of a replicate pair. Individual sample call rates are higher.

| Illumina 1,536 SNP GoldenGate OPA |         |             |                        |             |                        |
|-----------------------------------|---------|-------------|------------------------|-------------|------------------------|
| line                              | # pairs | ALCHEMY     |                        | BeadStudio  |                        |
|                                   |         | concordance | call rate <sup>1</sup> | concordance | call rate <sup>1</sup> |
| Nipponbare                        | 21      | 99.9%       | 97.9%                  | 98.5%       | 99.0%                  |
| 9311                              | 21      | 99.4%       | 94.9%                  | 96.6%       | 95.9%                  |
| NPx9311 (F1)                      | 15      | 99.4%       | 82.3%                  | 98.1%       | 99.1%                  |
| all others                        | 4       | 97.4%       | 90.1%                  | 95.4%       | 94.3%                  |
| average                           |         | 99.5%       | 96.8%                  | 97.6%       | 98.3%                  |

| Affymetrix 44K GeneChip |         |             |                        |             |                        |
|-------------------------|---------|-------------|------------------------|-------------|------------------------|
| line                    | # pairs | ALCHEMY     |                        | BRLMM-P     |                        |
|                         |         | concordance | call rate <sup>1</sup> | concordance | call rate <sup>1</sup> |
| Nipponbare              | 21      | 99.9%       | 94.9%                  | 95.8%       | 65.0%                  |
| 9311                    | 10      | 99.8%       | 89.0%                  | 91.6%       | 63.6%                  |
| NPx9311 (F1)            | 15      | 99.8%       | 84.9%                  | 94.8%       | 71.1%                  |
| all others              | 20      | 99.4%       | 91.4%                  | 94.4%       | 66.6%                  |
| average                 |         | 99.7%       | 94.1%                  | 94.2%       | 66.6%                  |

and analyzing them to determine the genotypes expected (presuming homozygosity) for these lines from the sequence data, we can compare these expected genotypes to ALCHEMY calls. On average, we find a high concordance (average 99.1%, call rate 96.1%) with some of the lines having lower concordance being those derived from distinct plant materials or having lower coverage depth in re-sequencing. Taken together, these analyses broadly validate ALCHEMY’s genotype calls across many different rice samples.

Next, we asked whether or not ALCHEMY was over-fit to these specific genotyping products. The vendor’s genotyping algorithms work well for human products and other supported products, but did not perform well “out-of-the-box” on our custom arrays and our samples as demonstrated above. In the interest of promoting the development

Table 3.3: ALCHEMY vs BRLMM-P on 270 human HapMap Phase II Samples. Accuracy refers to the agreement between genotype calls for the respective algorithm and HapMap Phase II published genotypes

|           | ALCHEMY | BRLMM-P |
|-----------|---------|---------|
| accuracy  | 99.41%  | 99.82%  |
| call rate | 98.64%  | 99.60%  |

of new genotyping products in more systems, we would like to have a genotyping algorithm that performs well across a broad range of vendors, products, systems, and sample sets, requiring little or no empirical fine-tuning to obtain high quality data. To assess ALCHEMY’s performance on a non-rice data set, we obtained the publicly available HapMap Phase II published genotypes and the Affymetrix Human 500K GeneChip .CEL files that were run on these same samples and ran ALCHEMY and BRLMM-P (Table 3.3). As expected, BRLMM-P performs very well as it has been developed and tuned for this data set. Although ALCHEMY does not perform as well as BRLMM-P on human HapMap samples, it still performs very well even without specific tuning or trial-and-error adjustments to improve accuracy or call rate. These results, taken together with the results above showing strong results on two very different technologies, suggests ALCHEMY is a generalized method with broad applications, especially for custom products where fine-tuned specialized algorithms are not available.

Finally, since custom products may require optimization of molecular techniques and protocols to obtain optimal results, we wanted to develop a method which produced accurate and usable results even if the total sample size was small. To demonstrate ALCHEMY’s ability to call small sample sizes, we ran ALCHEMY as well as the vendor’s software on a series of sample subsets and assessed the accuracy and call rates for our three reference samples: Nipponbare, 9311, and the Nipponbare X 9311 F1. Additionally, we ran each of the three reference samples alone to assess performance on a single sample. Because of the unfortunate discrepancy between our 9311 line and the



Table 3.4: ALCHEMY vs BRLMM-P on single samples and small sample subsets.

| # of samples       | ALCHEMY  |           | BRLMM-P  |           |
|--------------------|----------|-----------|----------|-----------|
|                    | accuracy | call rate | accuracy | call rate |
| Nipponbare alone   | 99.74%   | 89.37%    | 67.06%   | 88.48%    |
| 9311 alone         | 99.88%   | 90.93%    | 76.13%   | 90.70%    |
| NPx9311 (F1) alone | 84.64%   | 89.16%    | 89.67%   | 88.07%    |
| 3 (full trio)      | 98.08%   | 89.88%    | 81.31%   | 87.28%    |
| 6                  | 99.12%   | 92.07%    | 82.41%   | 85.07%    |
| 9                  | 99.46%   | 92.14%    | 84.05%   | 84.88%    |
| 12                 | 99.57%   | 92.16%    | 84.79%   | 82.69%    |

9311 line which was sequenced, we gauge “accuracy” on the basis of agreement with ALCHEMY calls for these samples on the full data set but restrict ourselves to SNPs which are consistent with Mendelian transmission to the F1 in the full data set. In table 3.4, we find that ALCHEMY quickly attains >99% accuracy with as few as 6 samples. Additionally, it performs very well on either homozygote sample alone, but poorly on the heterozygote sample alone. In contrast, BRLMM-P performs very poorly on any of the three reference samples alone and poorly on small samples sizes. BRLMM-P accuracy increases as sample size increases, as expected, but surprisingly, call rates decline. Similar results are observed with Illumina BeadStudio, except call rates do not decline with larger sample size (not shown).

### 3.6 Discussion

The design of ALCHEMY was motivated primarily by two concerns: (1) the poor performance of the vendor’s software on inbred sample sets and (2) the requirement for a large number of samples to be simultaneously analyzed to obtain accurate results. As mentioned previously, the two concerns are related, as the main reason many samples

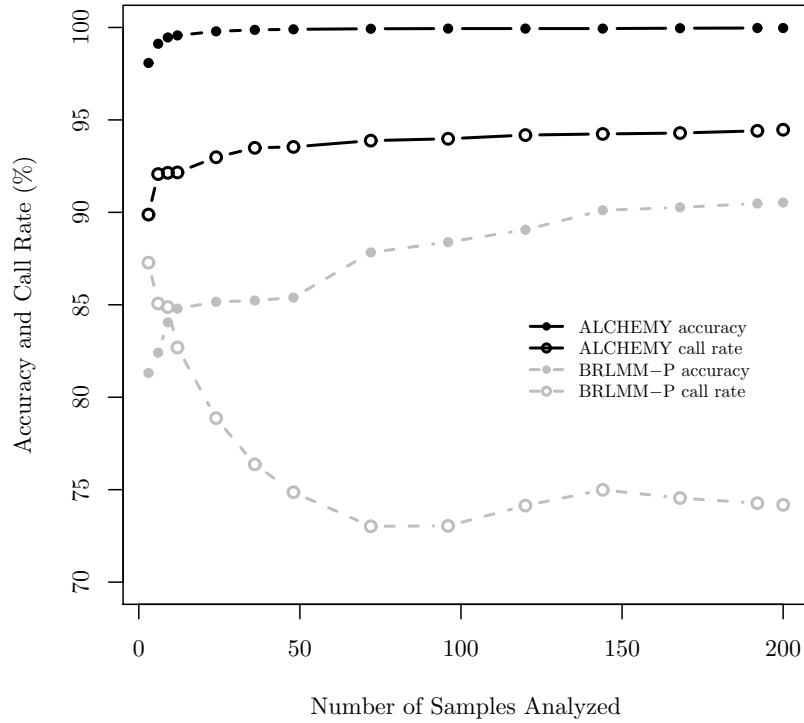


Figure 3.2: Effect of increasing number of samples which are simultaneously analyzed for ALCHEMY and BRLMM-P (Affymetrix 44K)

are required for clustering algorithms is to ensure that each genotype cluster is well represented allowing its location and boundaries to be well defined. Thus, if heterozygotes are rare or absent in the data due to inbreeding, the heterozygote cluster can not be reliably identified even if large numbers of inbred samples are used.

To address this, we have proposed a statistical model to describe the raw intensity data which is the basic observation of both Affymetrix and Illumina genotyping platforms. The model is capable of making an inference even if only a single sample is analyzed, but the parameters of the model are refined and optimized when several samples are available for simultaneous inference. In addressing concern (2), this approach

is shown to be highly successful, with ALCHEMY obtaining >99% accuracy with as few as 6 samples, and larger number of samples continuing to improve call rates. Additionally, the statistical treatment of the problem permits inbreeding to be explicitly considered and incorporated into the model in an appropriate way. Simultaneously estimating and optimizing the inbreeding coefficient on a per-sample basis allows both outbred and inbred samples to be analyzed simultaneously and improves both accuracy and call rates.

Although not studied extensively here, the posterior call probability produced by ALCHEMY as a quality metric can conceivably be used directly in downstream population genetic and quantitative genetic statistical analyses. For the rice data sets and the human HapMap data set, the signal-to-noise ratio of the intensity data is strong enough that for most SNPs there is little uncertainty in the genotype call. However, in noisier data, incorporating the probability of error estimated by ALCHEMY for genotype calls into statistical analyses may allow for more accurate population genetic inferences and improve both sensitivity and specificity of genome wide association studies.

The results presented here show that ALCHEMY's performance is superior to either vendor's standard software on the two rice genotyping products considered. Additionally, the strong performance on human HapMap data suggests ALCHEMY may work well on a wide range of products. We have also tested ALCHEMY on currently unpublished data from dogs in both Affymetrix and Illumina products and found a consistent high level of performance at or exceeding the levels reported here. While BRLMM-P out performs ALCHEMY on the HapMap samples, it is important to note that we did not attempt to tune or alter ALCHEMY for improved performance on HapMap as the

purpose was to test whether or not ALCHEMY is already over-fit to the rice genotyping arrays for which it was developed. There are several options to the program which may improve performance in specific applications. In practice, users will want to experiment with these options to obtain optimal results. Such options include the number of degrees of freedom for the bi-variate t-distribution, whether or not to estimate heterozygote signal means from the data or use a fixed proportion of estimated homozygote signal levels, specifying *a priori* allele frequencies, whether or not to conduct an EM search for optimal inbreeding coefficient values or use fixed values, and changing the fixed covariance matrices to reflect stronger or weaker covariances observed.

It should be possible to estimate the covariance matrices from the data in the same fashion that SNP-specific signal and noise means are estimated. However, in practice we found that correlation in signal intensities varies greatly between SNPs and accurate estimates require a large number of samples. Since a primary design goal was robust performance on small number of samples, we opted to fix the covariance matrices to values typical of our observed data which delivered good performance on empirical measures. Surprisingly, the same values we used for rice provided very good performance on the human HapMap samples which suggests that optimizing these parameters is not necessary to obtain >99% accuracy.

A novel feature of ALCHEMY is the explicit handling of inbreeding and the ability to simultaneously estimate both genotypes and inbreeding levels from the raw intensity data. The prior probabilities only hold for a single population however, and the presence of sub-population structure may result in over-estimation of inbreeding coefficients. However, for genotype calling in ALCHEMY, only the reduction in expected

heterozygosity is relevant, not whether it is due to inbreeding or population structure. In principle however, it is possible to extend ALCHEMY such that population structure, and inbreeding within sub-populations, is simultaneously estimated along with genotype calls, potentially improving the accuracy of both.

## **Availability**

ALCHEMY is written in C and developed and used under the GNU/Linux environment. It is available free of charge for both commercial and academic use under the terms of the GNU General Public License version 3. Source code and documentation is available at <http://alchemy.sourceforge.net/>. Source code is expected to compile and run on any GNU/Linux platform, Mac OS X, and Unix environments with the GNU C compiler and associated tools installed.

## **Acknowledgments**

DNA and plant material for the Nipponbare x 9311 F1 as well as parent lines was provided by Dr. Guo-Liang Wang. Dr. Brian Scheffler on behalf of the RiceCAP project (USDA/CSREES grant 2004-35317-14867) provided raw Illumina GenomeAnalyzer II reads for lines Cypress and LaGrue which were used in validating genotype calls on these lines.

## CHAPTER 4

# **ROBUST COMPOSITE LIKELIHOOD METHOD FOR GENOME WIDE SELECTION SCANS REVEALS A LARGELY INDEPENDENT SELECTIVE HISTORY OF *ORYZA SATIVA* SUBPOPULATIONS COMBINED WITH INTROGRESSIONS OF SELECTED ALLELES BETWEEN SUBPOPULATIONS**

**Abstract.** Since the availability of molecular population genetic data, many tests for selection have been proposed and applied to datasets of varying size, from single gene loci to entire genomes. Most of these tests are based on ad-hoc statistics deemed to be sensitive to selective events, and some on more rigorous mathematical theory, but all suffer from the fact that the null distribution of the statistic is unknown and very sensitive to non-selective parameters of the data such as recent fluctuations in population size, mutation rate or SNP density, and local recombination rate. In order to compare the extent of selective events harbored by distinct but related populations, the ability to accurately determine statistical significance is critical. Here, we present an extension of an existing composite likelihood method by introducing a permutation test to determine statistical significance and show this method is accurate and robust. The method is applied to 39 genomes from 4 subpopulations of *Oryza sativa* (domestic rice) and 8 genomes of *Oryza rufipogon* (common wild rice). We find that each of the 4 subpopulations harbors a substantial number of selective fixations that are largely independent and subpopulation specific. An analysis of haplotype sharing between subpopulations at selected loci revealed that many selective fixations, originating in a single subpopulation, are introgressed into other subpopulations, with the *aus* subpopulation being an exception in that it was neither the donor nor recipient for introgressions at selected loci. Furthermore, our results indicate that selected alleles from cultivated subpopulations are introgressed back into the wild progenitor *Oryza rufipogon*, or at least into accessions maintained by germplasm reserves, but not in reverse. These results suggest that domestication, as an ongoing process, is largely proceeding independently within subpopulations, but frequently influenced by improvements developed in other subpopulations.

## 4.1 Introduction

One of the strongest and well characterized predictions in the mathematical treatment of evolution and population genetics is the effect of a recent selective fixation on linked neutral variation. As a selected allele rapidly increases in frequency in a population, the genetic background on which that allele arose also increases in frequency unless recombination causes the selected allele to appear on different genetic backgrounds as well. Since recombination frequency increases with genetic distance from the selected site, proximal sites are likely to be fixed along with the selected allele. Likewise, variation at more distal sites may remain segregating after the selected allele has fixed, but allele frequencies increased or decreased from neutral expectations. The localized loss-of-diversity at and around a selected site which has gone to fixation is known as a “selective-sweep”, and the co-fixation of alleles at tightly linked loci is often referred to as “genetic hitch-hiking.”

A mathematical treatment of this hitch-hiking effect was first presented by Smith and Haigh (1974), who characterized the localized depression along a chromosome in expected heterozygosity due to selective fixation and demonstrated that this depression decreases as distance from the selected site increases. Other earlier work focused on the effects of selective fixation on the site frequency spectrum, but ignored the spatial pattern described by Smith & Haigh. This includes familiar summary statistics such as Tajima’s  $D$  (Tajima, 1989), Fu and Li’s  $D$  (Fu and Li, 1993), and Fay and Wu’s  $H$  (Fay and Wu, 2000). While these methods are easy to use, it is widely documented in the literature that they are prone to false positives under many common, non-equilibrium circumstances, such as recent changes in effective population size (Przeworski, 2002; Wakeley and Aliacar, 2001; Jensen *et al.*, 2005). Thus, these tests are better described

as tests of the constant-size neutral equilibrium model, as Tajima originally described his D statistic.

Kim and Stephan (2002) were the first to capture the spatial pattern of variation in the site frequency spectrum induced by a recent selective sweep. In their method, the probability distribution of allele frequencies at linked sites is specified as a function of distance to the selection target and the strength of selection. However, the frequencies of neutral alleles at different sites are not independent due to linkage disequilibrium and shared ancestry between sites. Neither these pairwise nor higher order correlations can be mathematically described, so Kim and Stephan present a “composite likelihood” formula which is simply a product of the likelihood at each individual site, treating them as statistically independent and ignoring the correlation between sites. Conveniently, maximizing this composite likelihood function also maximizes the true, unknown likelihood function, providing a maximum likelihood framework for estimating both the location of the selected site and the strength of selection. A neutral, null model composite likelihood can also be calculated and a test statistic constructed from the ratio of the composite likelihoods of the maximized sweep model and the neutral model.

Unfortunately, unlike true likelihood ratio statistics for nested models, the distribution of the composite likelihood ratio of Kim and Stephan under the null hypothesis (no selection) is unknown and can not be derived, due to the unmodelled correlation between sites. Thus, to determine statistical significance, extensive simulations of neutral population samples must be performed and an empirical null distribution tabulated.

Although Kim and Stephan provided a significant advance in methods to detect re-



cent selection, this method is also prone to false positives if a population has experienced recent changes in effective population size as well as other non-selective violations of the constant size neutral-equilibrium model. Jensen *et al.* (2005) considered these cases and proposed a goodness-of-fit test to distinguish rejections in the Kim and Stephan test due to non-equilibrium population history vs. a recent selective sweep. In this test, the Kim and Stephan method is used to provide the maximum likelihood estimates of the target (position) and strength of selection. Using these estimated parameters, data is simulated under a selective sweep model and the variation generated in these simulations is compared to the observed data. Using a summary statistic, the observed data is compared to the distribution of this statistic under many simulated selective sweeps, and the hypothesis of selection is rejected in favor of a non-selective violation of neutral-equilibrium if the observed data statistic is in the tails of the simulated distribution (ie, the observed data is atypical for a sweep simulated with the maximum-likelihood estimates of selection strength and position). For many of the cases Jensen *et al.* (2005) considered, this approach appears successful, but excessive false positives are still found in the case of extreme (99% reduction) bottlenecks and certain population substructure models.

Recently, the resequencing of entire genomes and the development of high-throughput genome-wide SNP arrays has provided an unprecedented opportunity to analyze the genome for targets of recent selection. While Kim and Stephan and Jensen *et al.* consider the analysis of a single locus ( $\approx 10$  Kb of contiguous sequence), studying the entire genome provides an obvious idea to control for population history. Population demographic events, such as fluctuations in population size, affect the entire genome equally in expectation, but a selective sweep has a localized effect. Therefore, if selection is sufficiently rare in the genome, it may be identified as regions of the genome

which are consistent with predictions of a selective sweep, but inconsistent with the overall pattern observed across the genome.

Nielsen *et al.* (2005) adapted the approach of Kim and Stephan for use with genome wide data, proposing to derive the “background site frequency spectrum” from the genome as a whole, and use this to derive both the basis for the expectations under the sweep model as well as for the neutral model in the composite likelihood framework. While Kim and Stephan consider full resequencing data where invariable sites are known, Nielsen considers SNP data where intervening sites have not been assayed and whether or not they are variable is unknown. Additionally, for high-throughput SNP genotyping data, the SNPs which are assayed represent a subset of the variable sites in the genome and are typically subject to an “ascertainment bias” (Nielsen, 2004; Clark *et al.*, 2005) in that these assayed SNPs must have first been discovered in a shallow complete re-sequencing sample (ascertainment sample). Because the ascertainment sample is typically much smaller than the genotyped sample, the frequency distribution of the genotyped SNPs is biased toward mid-frequency alleles which are more likely to be observed as variable in the ascertainment sample. This presents a serious problem for analyses of this kind, since composite likelihood under the sweep model is directly a function of allele frequencies.

Like Kim and Stephan, the null distribution of the Nielsen *et al.* composite likelihood ratio must be obtained through extensive simulations under a neutral model. It has been claimed in Nielsen *et al.* (2005) that the use of the genome-wide background frequency spectrum provides robustness to fluctuations in population size and that the effects of ascertainment bias may be modeled directly in the composite likelihood for-

mulae. Using the same methods, Williamson *et al.* (2007) further claim that the method is not sensitive to recombination hotspots and only weakly dependent on local variation in recombination rate in the genome. Although the majority of the mass of the null distribution computed under several different neutral models appears to overlap, as shown in these papers, we show here that when considering the upper tail of the null distributions, wide deviations are apparent. Due to the high multiplicity of statistical tests inherent in genome-wide scans, it is precisely the upper tail (eg, beyond  $p < 0.001$ ) that matters most. Thus, as before, the accuracy of p-values for statistical tests of selection depend strongly on the ability to accurately capture the non-selective aspects of the population history and the data in the null distribution simulations.

Many studies using these methods are aware of the inability to precisely give statistical significance to the results, so the authors guard against false positives by making very conservative assumptions. For example, Williamson *et al.* (2007) claim that the null distribution provided by assuming a constant size population is more extreme than more plausible demographies in human, and additionally reduce their estimates of recombination rates by a factor of 5 to guard conservatively against inaccuracy in these estimates. As a result, it is difficult to make any claims about the pervasiveness of selection across the genome, or compare between diverged populations to identify shared or subpopulation-specific selective events.

In search of a more robust composite-likelihood method to assess statistical significance while avoiding the need to make conservative assumptions that reduce power, we present here a simple, permutation-based method which disrupts the spatial pattern produced by a selective sweep but retains all salient features of the genomic background

such as pairwise and higher-order linkage disequilibrium, the location of SNPs in the genome, and the allele frequency distribution. We show that this method produces accurate p-values in the cases of recent population size changes and fluctuations in local recombination rates while retaining power to detect selection.

Using this new method, we analyze 39 resequenced genomes of domesticated rice *Oryza sativa* and 8 resequenced genomes of its wild progenitor *Oryza rufipogon*. The *Oryza sativa* sample represents 4 well recognized subpopulations of cultivated rice (Garris *et al.*, 2005), *temperate japonica* (11), *tropical japonica* (10), *indica* (7), and *aus* (11), which were analyzed for selective sweeps separately. As a domesticated plant, *Oryza sativa* may be expected to have experienced several selective sweeps in its recent history during domestication. Although studies have been performed with a large number of loci representative of the genome and have documented evidence that a domestication-associated bottleneck alone can not explain the patterns of variation observed (Caicedo *et al.*, 2007), the extent to which selection has impacted the genome of these cultivated rice subpopulations has not been studied comprehensively.

Here, we show that selection is common in the genome of each subpopulation and exceeds that found in its wild progenitor *Oryza rufipogon*, but a large number of selective sweeps are detected in this species as well. Surprisingly, subpopulations share few selection target predictions although the number shared in pairwise comparisons is statistically significant and not due to random co-occurrence. An analysis of haplotype sharing at common sweeps between pairwise comparisons did not find significant evidence that sweeps in these regions have been inherited from a common ancestral subpopulation. Selection at these loci may have occurred independently in these subpopu-

lations and their non-random co-locations due to common selective pressures targeting the same genes. Conversely, comparisons of haplotype sharing between subpopulations at loci where only one subpopulation had a predicted selective sweep reveals increased haplotype sharing in some cases indicating that selective events originating in one subpopulation may have been introgressed into another, but selection for recombinants during the introgressive process is distinct from the classic selective sweep model and likely would not be detected by the composite likelihood method.

We interpret these results as supporting the independent-origin (Caicedo *et al.*, 2007; Guo *et al.*, 2008) hypothesis of rice domestication due to the largely independent selective history found. Selective sweeps shared across multiple subpopulations by descent, a finding which might support the single-origin hypothesis (Gao and Innan, 2008; Vaughan *et al.*, 2008), were not found.

## 4.2 Methods

**Data** Resequencing data from Illumina GenomeAnalyzerIIx was obtained directly and from several collaborators. In simulation studies, we found that roughly 80% of the reference Nipponbare (*temperate japonica*) is accessible from paired-end reads at 86 bp read lengths and 7X coverage (average of the obtained data), and to qualify for this study sequence from an individual line must obtain 2X or greater coverage of at least 70% of the reference genome. This included 11 *temperate japonica* lines, 10 *tropical japonica* lines, 7 *indica* lines, 11 *aus* lines, and 8 *Oryza rufipogon* lines. Additionally, *Oryza meridionalis* was resequenced as an outgroup. 70.1% of SNPs were polarized

into ancestral and derived variants using data from *Oryza meridionalis*.

**Primary Sequence Analysis.** Resequencing data was converted into SNP data and haplotypes using a software package called PANATI (<http://panati.sourceforge.net/>). Details of PANATI are described on the website. Briefly, PANATI is a map-to-reference next-generation sequence analysis suite designed originally for SNP discovery for the purpose of designing fixed-array genotyping products. As part of the SNP discovery strategy, PANATI integrates data across several lines to both call SNPs and create haplotypes. To call SNPs in individual lines, at least 2X coverage at a site was required with 2 or more reads supporting a single non-reference allele. In our materials, all sequenced lines including the *Oryza rufipogon* samples have been extensively inbred and are essentially homozygous across the entire genome. Thus, the presence of a single read with the reference allele disqualifies a SNP call from an individual line. In addition to requiring two or more reads supporting a non-reference allele and no reads carrying a reference allele, the base quality score of at least one of the reads must exceed a threshold of 20. The lists of SNPs discovered in each line are then merged to create a master list of SNPs discovered across the entire sample. Additional SNPs are added to the master list if 2 or more lines carrying a single read with a non-reference allele (and no reference allele reads) with base quality score exceeding 20 are found. Using this master list, the PANATI output is then queried for the base observed in the resequencing data resulting in a haplotype for each line across these SNPs. Due to lack of complete coverage of the genome, some lines may have no data at a portion of SNP sites and the allelic state in the haplotype is coded as missing data in these cases. This procedure is also performed using an inbred *Oryza meridionalis* sample as an outgroup, except this sample is not allowed add SNPs to the master SNP list and instead is only queried for the alleles observed at each SNP on the master list. For 70.1% of the SNPs on the mas-

ter list, the allelic state in each line could be polarized as ancestral or derived using the outgroup. Using simulated data from the Nipponbare rice reference genome incorporating true SNPs, sequencing errors, and insertions and deletions, the false discovery rate of this procedure was estimated at 0.3%. In addition, we resequenced the Nipponbare *temperate japonica* line from which the high-quality reference genome sequence was derived and measured an empirical rate of false SNP calls at 0.03 SNP/Kb. As the plant material used for our resequencing is distinct from that used for the reference genome, many of these SNP calls may reflect true residual diversity in germplasm collections of inbred lines.

**SNP data.** In total, 12.9 M SNPs were discovered across the entire dataset, with 1.2 M segregating within *temperate japonica*<sup>1</sup>, 2.5 M in *tropical japonica*, 4.3 M in *indica*, 4.3 M in *aus*, and 7.1 M in *Oryza rufipogon*, corresponding to  $\theta_W$  (Watterson's estimate of the scaled mutation rate which is comparable across different sample sizes) of 0.0011, 0.0023, 0.0047, 0.0040, and 0.0074 respectively. This is consistent with previous work showing that *temperate japonica* is the least diverse and most bottlenecked *Oryza sativa* subpopulation, and *indica* and *aus* are much more diverse than either *japonica* subpopulation (Garris *et al.*, 2005; Caicedo *et al.*, 2007). As well, all cultivated populations are less diverse than their wild relative *Oryza rufipogon*. As invariant sites within subpopulations may be used in the sweep and null models, all sites were used in all populations regardless of whether they were segregating or not within subpopulation. This increases power to detect selection, but at a low computational cost as only sites known to be segregating in other populations are used as fixed sites. While the CLR statistic is sensitive to local variations in mutation rate and SNP density, the permutation test for significance described here is not as the SNP density and their locations relative to the selection tar-

---

<sup>1</sup>the reference temperate japonica line nipponbare was included for determining segregating sites in temperate japonica, and excluded for other subpopulations

get is preserved in the permuted data.

**Selection scan.** For the rice populations analyzed in this study, the genome was divided into 100 Kb segments, within which the selective sweep model and corresponding null “background” model was evaluated at 11 points corresponding to the endpoints of the interval and every 10 Kb in between. The CLR value for each successive two points was averaged and the 10 Kb interval containing the maximum was explored further at every 1 kb in between the endpoints. Assuming the likelihood surface within this range is fairly regular and smooth, this results in selecting the maximum CLR position along a 1 Kb grid with only 20 evaluations of the likelihood function. Each evaluation of the likelihood function is computationally expensive because the strength parameter must also be optimized for each position. Similar to the nested grid search for the maximum position, the log of the strength parameter  $\alpha = r/s \log(2N)$  (see Nielsen *et al.* (2005)) is also optimized on a 2-stage nested grid. Permutation as described was performed for up to 100,000 trials, with evaluations for a 100 Kb segment terminating once 20 trials had demonstrated a larger CLR value than the observed data. For positions that did not reach 20 by 100,000 trials, p-values were estimated using the first 10,000 permutation CLR values which are retained in memory. As described, a linear model is fit to the CLR value vs. the log of its quantile, using only CLR values with a quantile above 0.10 and below 0.95 to avoid the non-exponential lower tail as well as the the upper range where quantile estimates are inaccurate. In all cases,  $r^2$  of such regressions exceeded 0.99. As described, declaration of a selective sweep requires not only that the target with the 100 Kb window be statistically significant, but that the CLR likelihood surface decline to 50% or less of the candidate target CLR, or that the p-value in adjacent 100 Kb segments rise to at least 0.01 before the CLR rises again to another target. This additional criterion is required to account for the fact that a strong selective sweep will



likely cause several adjacent 100 Kb windows to be statistically significant even though only one sweep has occurred.

The program is designed to handle varying sample sizes across sites and a mixture of polarized (ancestral and derived allele distinguished) and unpolarized sites. The former case is handled by first estimating the background site frequency spectrum across the genome using only SNPs where complete data is available. Lower depth site frequency spectra are then estimated from this using the hypergeometric distribution which describes drawing a subsample without replacement. These background site frequency spectra are then used to compute tables of site frequency spectra of corresponding depth under the sweep model for a range of distances from the target site and strengths of selection using the formulas and equations presented in Nielsen *et al.* (2005). Since the expectation of derived allele frequency is a smooth function of selection strength and distance, spline functions are fit to these sweep model site frequency spectrum tables and used in the computation of the sweep model and maximum-likelihood estimation of strength and position parameters. For unpolarized SNPs, the probability of observing a SNP with allele frequency  $d < n - d$  is simply the sum of the probability of observing  $d$  and  $n - d$  in the polarized site frequency spectrum, for either the sweep model or null genome-background model case.

**Shared sweeps and pairwise comparisons.** From the results of individual subpopulation scans, selective sweep targets were deemed potentially in common if the target predictions were within 100 Kb of each other, as this is approximately the resolution of the individual scans. It was not required that common sweep targets be within the same 100 Kb block of the reference genome. The probability of observing  $n$  or more

sweeps within 100 Kb of each other, given the total number of selective sweeps in each subpopulation and the number of 100 Kb windows in the genome was computed using the hypergeometric distribution.

**Haplotype sharing.** As described in the results section, the sHH statistic (shared haplotype homozygosity) was used to determine the relative extent of haplotype sharing between subpopulations. The analog of extended haplotype homozygosity (EHH, Sabeti *et al.* (2002)) was computed but between subpopulations, where instead of both chromosomes being drawn from the same subpopulation, one chromosome was drawn from each subpopulation and putative identical-by-descent (IBD) tracts (all allele states identical for a run of SNPs) for all pairwise comparisons identified. sHH is defined as the average of EHH at successive pairs of SNP sites extending out from a central site, in each direction, multiplied by the distance from the distal SNP in each pair to the central site (trapezoid approximation to integration). Thus, sHH is analogous to iHS (Voight *et al.*, 2006) except between subpopulations instead of within subpopulations, and the central site is defined by a equally spaced grid along the chromosome rather than by SNP locations. The sHH statistic will generate higher values when multiple chromosomes in each subpopulation share putative IBD tracts rather than when only a few chromosomes show long tracts of IBD. More importantly, sHH will be higher when a large number of chromosomes from each subpopulation shares a haplotype rather than when one population has large tracts of IBD across many or all chromosomes within the subpopulation, as expected at a selective sweep, but another subpopulation only contains a few chromosomes that are IBD with those of the first.

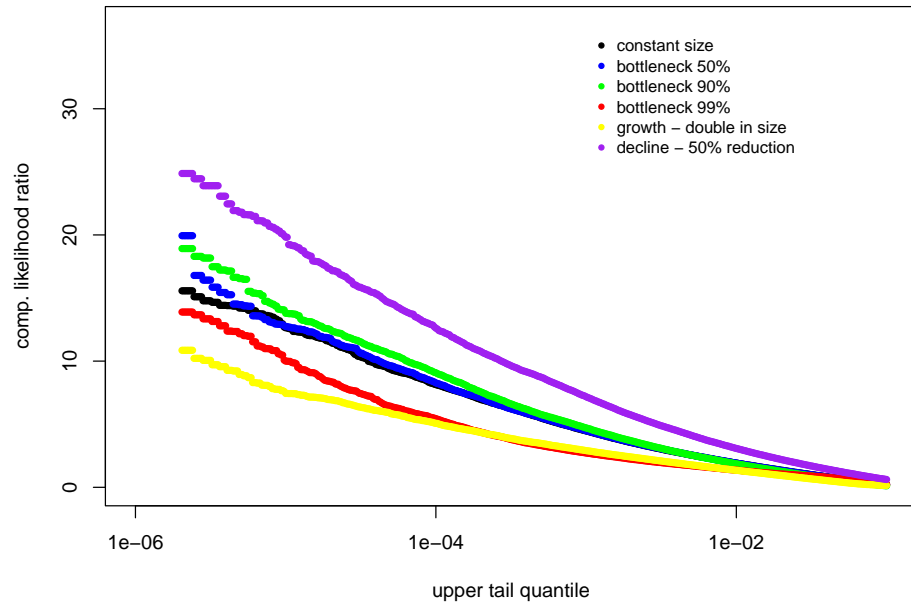
### 4.3 Results

Throughout the history of work on methods to detect recent selection using the distribution and frequency of polymorphic sites in a population sample, nearly all approaches thus far have utilized simulations of neutral data to assess statistical significance. This approach is simple in concept: simulate population samples under some assumed population model without selection, matching population and sample properties such as sample size, number of SNPs (or scaled mutation rate), and scaled recombination rate to that of the observed data. For a large number of such simulated samples, the statistic(s) used in the method is calculated for each neutral sample and an empirical distribution of the statistic under neutrality is obtained. Usually, it is the upper tail of this empirical distribution that will define what is considered statistically significant in the observed data. This procedure works for site frequency spectrum summary statistics such as Tajima's D (Tajima, 1989) and Fay & Wu's H (Fay and Wu, 2000), as well as the composite likelihood (CL) methods of Kim and Stephan (2002) and Nielsen *et al.* (2005). Conceivably, this procedure could work for haplotype based ad-hoc statistics (not based on a mathematical model) such as EHH (Sabeti *et al.*, 2002) and iHS (Voight *et al.*, 2006), although for these cases the accuracy of statistical significance will depend greatly on the estimate of the local recombination rate. Similarly, for CL methods and SFS summary statistic methods, the accuracy of p-values obtained by this approach greatly depends on extent to which the neutral model simulation accurately captures features of the population and sample that also affect the test statistic.

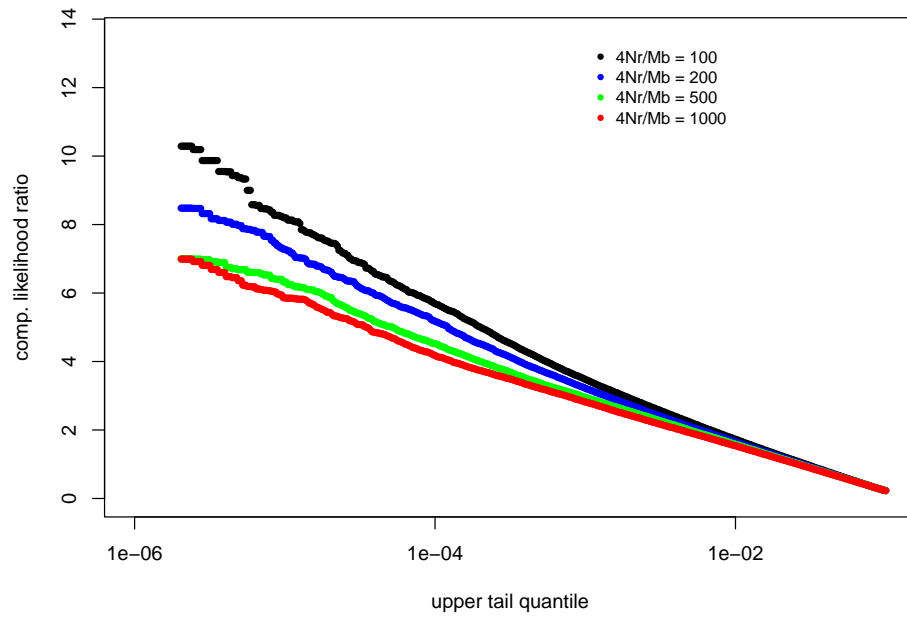
Nielsen *et al.* (2005) showed that the vast majority of the composite likelihood ratio (CLR) distribution mass between different human population demographic history scenarios is very similar, implying that simulation under an incorrect demographic model

would produce nearly the same CLR null distribution as the correct, unknown model. Since the large number of tests conducted in a genome scan will require higher levels of significance to be obtained in order to account for the multiplicity of selection hypotheses, we asked whether or not the upper tails of the CLR distribution were similar under several demographic scenarios. In Figure 4.1(a), we show that at significance levels suitable for genome-wide scans (eg,  $p < 0.0001$ ), the critical value of the CLR null distribution differs strongly between different neutral population size change models. Williamson *et al.* (2007) claimed that the constant size model provides the most conservative null distribution, yet we see here that a population decline or a strong population bottleneck inflates the CLR statistic and assuming a constant size model for the null distribution would produce an anti-conservative test in these cases.

We also show in Figure 4.1(b) that the population (scaled) recombination rate also affects the CLR distribution. Thus, if the null distribution of the CLR statistic is obtained by simulation, these simulations require an accurate estimate of the recombination rate in the region studied, and perhaps several different simulations need to be performed with varying recombination rates and the appropriate null distribution matched to the observed data based on estimates of local recombination rate in the genome. A null distribution obtained from a single recombination rate assumption will result in a conservative or anti-conservative test for selection depending on the true local recombination rate which varies across the genome. Recognizing this, Williamson *et al.* (2007) use recombination rate estimates from the Human HapMap (Consortium, 2003), but reduce these estimates by a factor of 5 in their simulations to be conservative. This strategy reduces false positives but at an unknown cost in power.



(a) CLR distribution under various simulated neutral demographic models



(b) CLR distribution under different simulated recombination rates

Figure 4.1: Tail of the CLR distribution under different neutral demographic scenarios and different recombination rates. On log scale, it is readily seen that at significance levels suitable for genome-wide scans, the CLR distribution is highly dependent on these parameters.

In our search for a new method to assess significance, we turn our attention to the attractive idea of an internal control provided by the genome background as originally proposed by Nielsen *et al.* (2005), and we note that the salient feature of the selective sweep model described by Smith and Haigh (1974) on which the CL methods are based is the accounting for the position of variable sites with respect to an assumed target of selection. The valley of heterozygosity predicted by the classic selective sweep model may be completely abolished by the presence of only a few variable sites at neutral frequencies. Thus, we asked whether or not a permutation test could be constructed, by reordering the observed allele frequencies at SNPs to disrupt the spatial pattern expected under a sweep model, but preserving all other aspects of the data such as the genome-wide site frequency spectrum, density and position of SNPs, and short-range linkage-disequilibrium.

Permuting all SNPs uniformly and independently does not produce a valid significance test as this always results in a lower CLR distribution than that observed with neutral data (not shown). This is because linkage disequilibrium is present in the observed data ordering, but destroyed if the SNPs are completely permuted. The null model composite likelihood assumes pairwise and higher-order independence and can not capture this linkage disequilibrium. The sweep model however, while also assuming pairwise and higher order independence, with SNP frequency expectations being a function of distance from the target site, can weakly capture linkage disequilibrium to the extent that such disequilibrium results in correlated allele frequencies. As such, the CLR statistic on the observed data ordering is always higher since the sweep model has a higher likelihood due to linkage disequilibrium alone.

Considering this, we asked whether or not permuting blocks of SNPs could preserve the short range linkage disequilibrium found in neutrally evolving regions, but disrupt the spatial pattern of a selective sweep and the long range linkage disequilibrium generated by the rapid rise of the selected allele. To do this, we tested a permutation protocol where random size blocks of SNPs were exchanged, with the block size drawn from an exponential distribution with mean rate equal to typical length of significant linkage-disequilibrium. To our surprise, this works remarkably well, under a wide variety of circumstances. Shown in Figure 4.2 are quantile-quantile plots comparing observed p-value distributions with the expected uniform distribution under neutrality, for a variety of conditions which we showed previously to have marked impact on the CLR distribution. In all cases shown, the permutation test recovers accurate p-values well out into the upper tail.

More formally, the permutation-test works as follows: first, a genome-wide scan on the observed data using the methods of Nielsen *et al.* (2005) is performed and observed values of the CLR statistic recorded. Following, the observed derived allele frequencies at SNPs are permuted across the genome, in exponential sized random blocks with a rate parameter such that blocks of linkage disequilibrium typical of the genome are preserved (permuted together as a block), while longer range spatial patterns of a selective sweep are eliminated. Note that the position of SNPs are preserved, only the frequency and sample depths are permuted. Since SNP density and spacing has a critical effect on CLR statistic, preserving this aspect of the original data is unique to this approach and very difficult to do in population sample simulation programs such as *ms* (Hudson, 2002). If the CLR statistic on the permuted data is larger than that of the observed data at the corresponding position in the genome, a counter  $p_i$  for that position is incremented. The original data ordering is then restored and a new random permutation computed. It

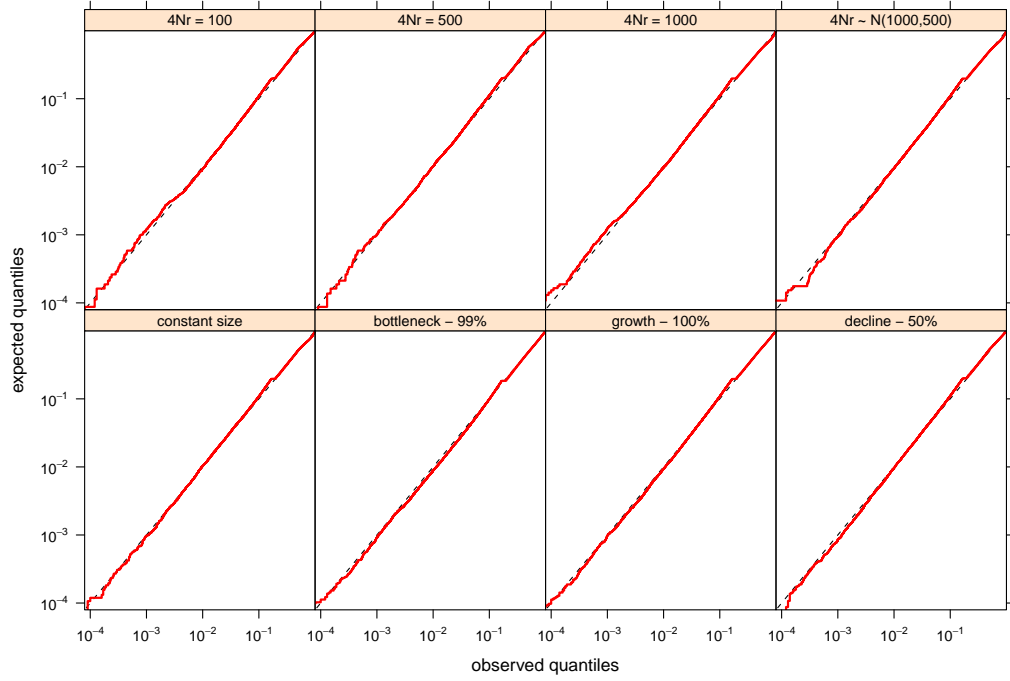


Figure 4.2: Quantile-quantile plots of p-values obtained by the proposed permutation method vs. expected p-values, for simulated neutral data generated under a variety of demographic scenarios and recombination rates. Note that axes are on log scale.

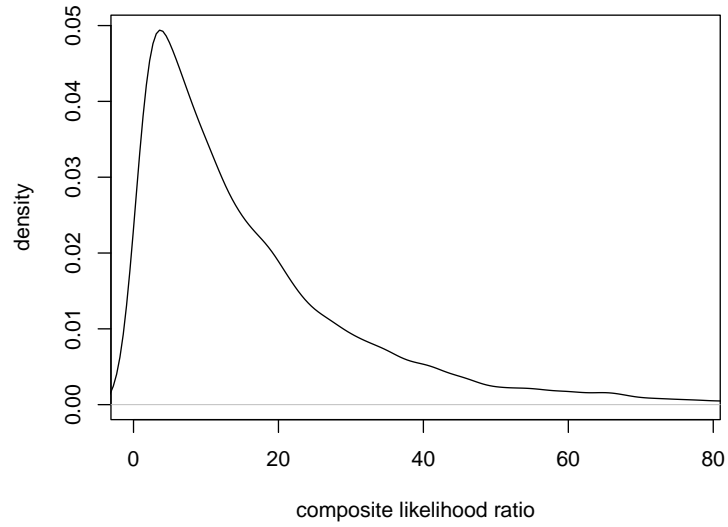
is important to restore the original data after each permutation or otherwise successive permutations will eventually completely disrupt the background linkage-disequilibrium typical of the genome. This procedure is repeated  $N$  times, and the p-value for position  $i$  in the genome estimated as  $p_i/(N - 1)$ . As this procedure is computationally expensive, we introduce an optimization where a locus in the genome is no longer evaluated once 20 permutations has produced a larger CLR value than the observed data. The p-value is then estimated as  $19/(N_i - 1)$  where  $N_i$  is the number of permutations of the whole genome performed when the 20th permutation exceeding the observed CLR value at position  $i$  was obtained (Haldane, 1945).

For the rice genome using the parameters described here (see methods), it is com-

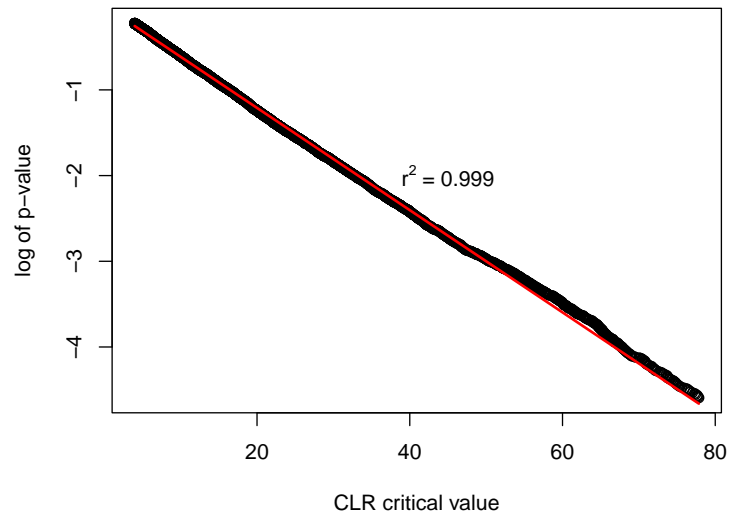


putationally tractable to perform this procedure up to 100,000 permutations, yielding the ability to distinguish sweeps of 0.0001 vs 0.00001 significance levels. To increase resolution to  $p < 10^{-6}$  using permutation alone requires 10 times the computation time, and  $p < 10^{-7}$  requiring 100 times more computational power. We note in Figure 4.3(a) that the CLR null distribution, as tallied by the permutation procedure, shows an exponential decay in its tail and appears as a straight line (Figure 4.3(b)) when the CLR statistic is plotted against the log of the quantile (log of the p-value). It is valuable to know which sweeps in the genome are the most significant, and in general the higher the CLR value the stronger the evidence for a sweep. However, CLR values at different loci in the genome are not directly comparable as the CLR statistic is dependent directly on SNP density. Thus, we use a linear regression of the log of the CLR distribution quantile versus the CLR value itself to extrapolate p-values beyond the range afforded by permutation alone (see methods for exact details). In all cases for the results presented here, the  $r^2$  of the linear regression exceeded 0.99 which indicates that some confidence in p-value extrapolation may be expected for several log units down from  $10^{-4}$  and is likely a better indicator of the most significant sweeps in the genome than the CLR value itself.

The sweep model used here has two parameters which are optimized to obtain the maximum (composite) likelihood: the position or target of selection, and the strength of selection. In order to define a fixed set of *a priori* hypotheses we divide the 373 Mb genome into 100 Kb segments, and hypothesize that each 100 Kb segment contains one selective sweep. In scanning the genome, we analyze each 100 Kb segment in turn and allow the position within that segment to be optimized on a 1 Kb grid within the segment, and likewise the strength of selection parameter is numerically optimized for each position on the 1 Kb grid. The position within the 100 Kb segment with the highest CLR becomes the predicted target of selection for that segment. Similarly, we repeat



(a) Typical CLR distribution



(b) Corresponding Log-linear model

Figure 4.3: Representative of typical CLR distributions, up to a scaling factor, the tail appears to follow an exponential decay. In panel B, a linear regression between the log of p-values and the corresponding critical value of the CLR distribution as determined by permutation is an excellent fit to the data and useful for extrapolating higher significance levels beyond that which is computationally feasible with permutation alone.

this procedure in 100 Kb segments for the permutation test, again allowing the position of the selection target to be optimized within the 100 Kb segment on a 1 Kb grid. The maximum CLR position in the permuted data need not correspond to the maximum CLR position of the observed data, as we are testing the hypothesis that the 100 Kb segment contains a selective sweep, not the *a posteriori* hypothesis that the maximum CLR position within the segment of the observed data is itself a sweep. In the case the null hypothesis (no selection) is true, this results in a valid test whereas testing only the position of the observed data CLR maximum within each 100 Kb segment biases in favor of the observed data containing a sweep. This procedure also clearly defines the number of hypotheses and thus the multiplicity burden of the genome-wide scan.

While it is advantageous to define a small segment size to obtain maximum resolution to detect closely spaced selective events, it is typically the case that a strong selective sweep will impact a large enough region in the genome that several adjacent segments will appear statistically significant although only one of them will contain the true selection target. Thus, while using statistical methods to determine if each individual 100 Kb segment experienced recent selection, we use a biologically motivated interpretation of these results to determine sweep calls. In order for the target position within a 100 Kb segment to be declared a sweep, we require that the CLR statistic drop to one-half the value of the candidate target before rising to a value exceeding the target, in both directions, or that the p-value for adjacent windows rises above 0.01 before exceeding the significance of the candidate target. Table 4.1 shows the number of statistically significant ( $p < 0.00001335$ ,  $\alpha = 0.05$  under Bonferroni correction for 3744 multiple tests) sweeps in each subpopulation that satisfy this criterion. Approximately 5% of the genome is implicated for most cultivated subpopulations. While this dataset represents the most comprehensive genome-wide dataset yet in rice, the depth of each

Table 4.1: Number of sweeps called in each subpopulation, subject to the rule that the CLR likelihood surface must drop below 50% of a candidate peak before rising again above it. The number of 100 Kb window statistically significant is much higher than the number of called sweeps.

| subpopulation      | number of targets | % of genome |
|--------------------|-------------------|-------------|
| temperate japonica | 196               | 5.2%        |
| aus                | 220               | 5.9%        |
| indica             | 171               | 4.6%        |
| tropical japonica  | 203               | 5.4%        |
| rufipogon          | 110               | 2.9%        |

subpopulation sample is barely sufficient to conduct such an analysis and this study may be regarded as under-powered. Indeed, the lowest number of sweeps in a cultivated subpopulation is seen in *indica* which is also the smallest sample depth.

A graphical view of the selective sweep map of *Oryza sativa* and *Oryza rufipogon* is shown in Figure 4.4. As can be readily seen, there is little concordance in selective target predictions between subpopulations. In pairwise comparisons however, some selective sweep targets are in common. We defined as selective sweep target in two distinct subpopulations as being in common if the target prediction was within 100 Kb of each other, although they need not be in the same 100 Kb block used in the genome scan. Table 4.3 shows the number of sweeps in common between subpopulation pairs. While the number of shared sweeps are small, between cultivated subpopulations it is highly statistically significant and not likely due to random co-occurrence of sweeps in the genome.

We then asked whether or not subpopulations share haplotypes to a significant extent at sweep loci vs. non-sweep loci. If a selective sweep occurring early in domestication was inherited into multiple now differentiated subpopulations, we'd expect haplotype

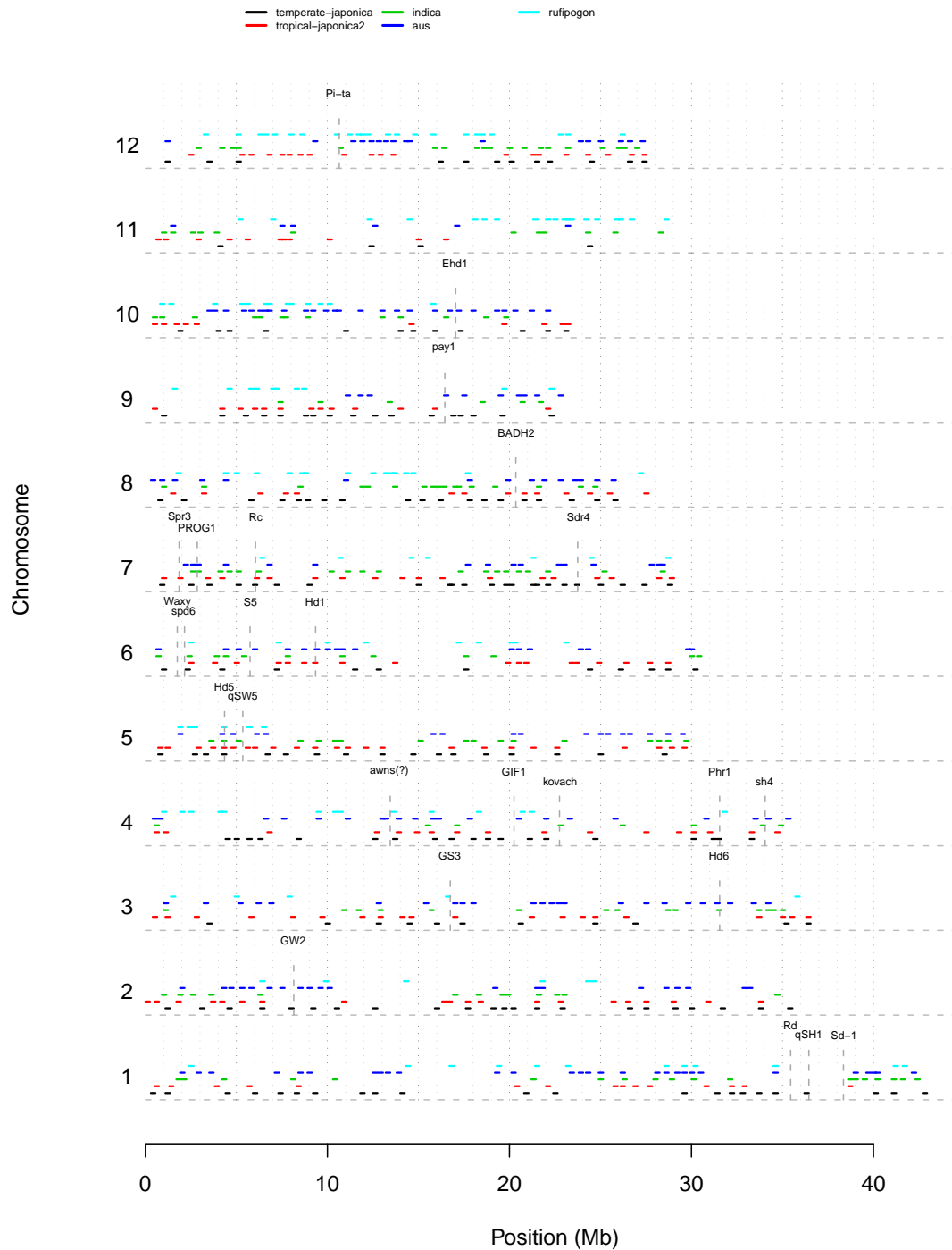


Figure 4.4: Graphical view of selective sweeps across the *Oryza sativa* and *Oryza rufipogon* genome, by subpopulation. Previously identified and characterized genes thought to be important in domestication are indicated by text centered at the gene's location and vertical dashed lines to indicate the precise position

sharing at common-by-descent sweep loci to be higher than that typical of genome-wide comparisons, simply as a result of reduced diversity inherited into both subpopulations at these loci. Conversely, if selection in each subpopulation occurred after these subpopulations diverged genetically, coincidental selection at the same loci, even if the same phenotypes were being selected, could potentially drive different haplotypes to high frequency or fixation, decreasing the level of haplotype sharing between subpopulation. Accordingly, we define an ad-hoc statistic to measure this called “sHH” (shared haplotype homozygosity), inspired by “iHS” introduced in Voight *et al.* (2006) but adapted for between subpopulation comparisons as opposed to within subpopulations (see methods for details). sHH is defined for any base in the genome and measures the level and extent of haplotype sharing between subpopulations extending in both directions away from a central base and scores higher values when a large number of chromosomes from each subpopulation are putatively identical-by-descent (IBD, approximated here as pairwise identical at all variable loci) over long distances.

This method allows us to ask an interesting question which is not captured well by the classic selective sweep model. In modern breeding practices, and perhaps as well in ancient practices, advantageous traits from one subpopulation are introgressed into another, with the goal of obtaining the advantageous trait while preserving all other traits of the recipient background. While this is a selective process, it is quite different from the classic selective sweep model which our primary analysis captures. During introgression, a breeder selects for *recombinants* whereas the classic selective sweep model specifies only selection on the desired trait and recombination with other backgrounds is stochastic and not favored over any other background carrying the selected allele. While the population dynamics of selection during repeated introgression into a majority of breeding materials in a subpopulation have not been studied, it is clear that

selection for recombinants would not result in a smooth valley of heterozygosity characteristic of the classic sweep model, and instead would result in a sharp transition from the selected region which is IBD with its donor, to genome-wide levels of variability and differentiation between subpopulations. Thus, we do not expect our primary selective sweep analysis to have sufficient power to detect selection in subpopulations where the selected target has been introgressed. If such introgression has occurred and is pervasive however, the sHH statistic at sweep loci should show higher values, on average, than other non-selected loci in the genome. Furthermore, by comparing sHH at selected loci where one subpopulation alone is used to define selected loci and the putative recipient population does not have a corresponding sweep prediction at the same location, we may be able to tell which way introgression between subpopulations occurred as well as the subpopulation in which the selected allele originated.

Table 4.2 shows the results of comparing the maximum value of sHH across a 100 Kb window centered on the “donor” subpopulation’s selective sweep targets, vs. the distribution of the maximum of sHH within all 100 Kb windows not implicated in selection in either the donor or “recipient” subpopulations. From this table, we draw several conclusions: *temperate japonica* and *indica* appear to have received a significant number of introgressions originating from loci detected as selective sweeps in *tropical japonica*, due to the significantly higher extent of haplotype sharing between these subpopulations and *tropical japonica* at classic selective sweep loci detected only in *tropical japonica*. Likewise, *tropical japonica* and *temperate japonica* appear to have received a significant number of introgressions from selective sweeps originating in *indica*, but the reverse is not true for *temperate japonica* sweeps introgressed into *indica*. The *aus* subpopulation appears to be the most independent cultivated subpopulation with the mean sHH value being lower at selective sweeps detected in *aus*, lower between *aus* and *temperate japonica*.

Table 4.2: Shared haplotype homozygosity (sHH) at sites where a “donor” subpopulation experienced a selective sweep and a potential introgression recipient population did not have a corresponding selective sweep called within 100 Kb. The sHH means shown are for the maximum sHH value across 100Kb intervals, as described in the text. Negative values of student’s  $t$  statistic indicate haplotypes are more diverse between subpopulations at the donor population’s selective sweep regions, positive values indicate greater haplotype sharing at these loci potentially indicating that putative selective events originating in the donor population may have been introgressed into the recipient population but are not detectable in the recipient as classic selective sweeps

| Subpopulation              | sHH mean<br>(sweep regions) | sHH mean<br>(non-sweep regions) | $t$    | p-value<br>(two-tailed) |
|----------------------------|-----------------------------|---------------------------------|--------|-------------------------|
| temperate-japonica (donor) |                             |                                 |        |                         |
| tropical-japonica          | 1.194                       | 1.098                           | 2.225  | 0.0261                  |
| indica                     | 1.469                       | 1.382                           | 1.950  | 0.0512                  |
| aus                        | 1.270                       | 1.315                           | -0.955 | 0.3395                  |
| rufipogon                  | 1.647                       | 1.600                           | 0.991  | 0.3216                  |
| tropical-japonica (donor)  |                             |                                 |        |                         |
| temperate-japonica         | 1.262                       | 1.089                           | 3.964  | 0.0001                  |
| indica                     | 1.535                       | 1.414                           | 2.825  | 0.0047                  |
| aus                        | 1.391                       | 1.368                           | 0.430  | 0.6669                  |
| rufipogon                  | 1.672                       | 1.659                           | 0.265  | 0.7913                  |
| indica (donor)             |                             |                                 |        |                         |
| temperate-japonica         | 1.532                       | 1.371                           | 2.573  | 0.0101                  |
| tropical-japonica          | 1.598                       | 1.408                           | 3.472  | 0.0005                  |
| aus                        | 1.673                       | 1.561                           | 1.899  | 0.0576                  |
| rufipogon                  | 1.731                       | 1.623                           | 2.275  | 0.0229                  |
| aus (donor)                |                             |                                 |        |                         |
| temperate-japonica         | 1.225                       | 1.322                           | -1.939 | 0.0526                  |
| tropical-japonica          | 1.370                       | 1.378                           | -0.165 | 0.8692                  |
| indica                     | 1.570                       | 1.579                           | -0.212 | 0.8325                  |
| rufipogon                  | 1.806                       | 1.800                           | 0.151  | 0.8803                  |
| rufipogon (donor)          |                             |                                 |        |                         |
| temperate-japonica         | 1.650                       | 1.606                           | 0.540  | 0.5889                  |
| tropical-japonica          | 1.689                       | 1.664                           | 0.283  | 0.7773                  |
| indica                     | 1.654                       | 1.639                           | 0.206  | 0.8367                  |
| aus                        | 1.955                       | 1.799                           | 2.014  | 0.0440                  |

*ica* at *temperate japonica* sweeps, and higher but not significant for *tropical japonica* and *indica* sweeps. No comparison with *rufipogon* was significant, suggesting that on average, selective sweeps in cultivated and wild populations have not been extensively



introgressed between them. We hasten to point out however that by comparing mean levels of haplotype sharing, we can not exclude the possibility that any individual sweep locus in any subpopulation has not been introgressed into any other subpopulation, and this analysis only reveals which populations, and in which directions, have seen extensive introgression of selected loci.

Next, we asked whether haplotype sharing at sweep targets in common between subpopulations was higher or lower than genome averages at non-selected loci. In Table 4.3, we see in the fourth column that the significance of seeing the number of shared targets (target predictions into two subpopulations within 100 Kb of each other) is highly significant for all cultivated pairwise comparisons compared to the expectation of random co-occurrence. Interestingly, haplotype sharing is significantly increased at common sweep loci for *temperate japonica* and *tropical japonica*, *tropical japonica* and *indica*, and *indica* and *aus*, but not significant in any other pairwise comparison. This suggests that some selective events occurred once and was inherited into present subpopulations from an ancestral population, but the pattern between subpopulations is difficult to interpret. The lack of putative common-by-descent sweeps between *temperate japonica* and *indica* and *temperate japonica* and *aus* suggests that *temperate japonica* did not share a common domesticated ancestor with either *indica* or *aus*. A possible interpretation of the putative common-by-descent sweep loci between *indica* and *aus*, but the lack of introgression, is that these two extant subpopulations shared a partially domesticated ancestral “*proto-indica*” population but *aus* was differentiated or isolated from what became *indica* very early on. Likewise, *temperate japonica* and *tropical japonica* may both be derived from a separate domestication of a “*proto-japonica*” population, but present day *tropical japonica* is now a mix of *proto-indica* and *proto-japonica* early domesticates, showing inheritance of selective sweeps that occurred during domestication

in both. We note however that we can not exclude the possibility that these “common” sweeps are actually still just introgressions that the composite-likelihood analysis managed to detect as a selective sweep even though the classic selective sweep model is poorly suited to detecting the selection of introgressions.

Table 4.3: Pairwise comparison of subpopulations for selective sweeps potentially in common between them. Between all cultivated subpopulation pairs, the number of sweeps detected within 100 Kb of each other is highly significant and not likely to be a random co-occurrence. The elevated haplotype sharing seen in some pairwise comparisons of cultivated populations may be indicative of selective events that were inherited from a common ancestral population and suggests possible relationships between these present-day populations and early domesticates.

| Subpopulations     |                   | targets<br>shared | p-value<br>(coincidence) | sHH mean<br>(sweep<br>regions) | sHH mean<br>(non-sweep<br>regions) | <i>t</i> | p-value<br>(two-tailed) |
|--------------------|-------------------|-------------------|--------------------------|--------------------------------|------------------------------------|----------|-------------------------|
| temperate japonica | tropical japonica | 41                | 0.0000                   | 1.409                          | 1.098                              | 4.618    | 0.0000                  |
| temperate japonica | indica            | 23                | 0.0000                   | 1.450                          | 1.382                              | 0.468    | 0.6397                  |
| temperate japonica | aus               | 24                | 0.0004                   | 1.328                          | 1.315                              | 0.078    | 0.9379                  |
| temperate japonica | rufipogon         | 8                 | 0.0514                   | 1.298                          | 1.600                              | -1.686   | 0.0932                  |
| tropical japonica  | indica            | 27                | 0.0000                   | 2.079                          | 1.414                              | 5.490    | 0.0000                  |
| tropical japonica  | aus               | 30                | 0.0000                   | 1.420                          | 1.368                              | 0.396    | 0.6919                  |
| tropical japonica  | rufipogon         | 6                 | 0.2558                   | 2.115                          | 1.659                              | 2.319    | 0.0219                  |
| indica             | aus               | 34                | 0.0000                   | 1.932                          | 1.561                              | 3.198    | 0.0014                  |
| indica             | rufipogon         | 9                 | 0.0092                   | 1.853                          | 1.623                              | 1.023    | 0.3080                  |
| aus                | rufipogon         | 13                | 0.0006                   | 2.000                          | 1.800                              | 0.826    | 0.4099                  |

## 4.4 Discussion

Research into tests of selection and their applications to genetic data has a long and detailed history and will likely continue. The permutation test presented here is another step forward towards statistically valid tests of selection in population genetic data, but should be viewed as an extension of the work of Nielsen *et al.* (2005). Previous methods have been shown to be highly sensitive to the effects of non-equilibrium demography such as population bottlenecks or growth. Simulating under the appropriate non-equilibrium but neutral model in theory should adequately account for these effects, but leave as a problem the estimation of these critical parameters. If simple, estimable bottleneck or growth models are not sufficient to describe variation at neutral loci, the only possibility is selecting a tractable model that is more conservative than the unknown true neutral model. Williamson *et al.* (2007) noted that the simulated scaled recombination rate affects the CLR null distribution and opted to use a conservative low recombination rate to guard against false positives. Another solution for improved power would be to have accurate estimates of local recombination rate across the genome and generate several null distributions for a selected range of recombination rates, but again this places the burden of accuracy on the estimation of local recombination rates, and also in addition to that of population demographic parameters.

The permutation procedure is an eloquent solution to all of these concerns, as nearly every aspect that can be learned from the data from one inference procedure or another, is internally maintained. Unique to this method is the preservation of SNP density and the position of SNP locations relative to an assumed target site. In addition to recombination rate and population demographic parameters, this too directly affects the null distribution of the CLR statistic and is difficult to match with simulated neutral data.

Using simulation, the best that can be done is to average over many different SNP position configurations, possibly at the loss of power.

The permutation procedure produces location specific null distributions and accurate p-values up to the limit of feasible computation. Luckily, it is readily seen that the CLR distribution is quite regular and always follows an exponential tail, just scaled differently at individual sites. In our procedure, we extrapolate more significant p-values by performing a linear regression between the CLR values produced by permutation and the log of their quantiles. So long as the lower tail (bottom 10% here) is ignored, the fit of the linear regression is nearly perfect and we might trust that p-values extrapolated several log units past that which can be estimated directly by permutation are accurate. This is of practical value in that it means that accurate p-values down to  $10^{-8}$  may be obtained with only  $10^4$  computational cost.

Here, we apply this method to populations of *Oryza sativa* and *Oryza rufipogon*, the former of which might be expected to harbor substantial amounts of selective fixations and for which the ability to quantify the number of those fixations is important. Not surprisingly, we find that all cultivated rice subpopulations have a large number of selective sweep predictions. In contrast however, it is striking that the vast majority of selective sweeps appear to be subpopulation specific with little concordance across subpopulations. This alone indicates that the majority of the domestication process, if viewed as an on-going process, has occurred independently in these subpopulations.

The number of “common sweeps” however was highly significant for all pairwise comparisons of cultivated subpopulations, suggesting that perhaps a few common

sweeps from an original domestication event have been inherited across two or more extant subpopulations. To investigate this, we analyzed the extent of shared haplotypes between subpopulations at common sweep loci and found that certain pairs of subpopulations show significantly elevated haplotype sharing, raising the possibility that these selective sweeps occurred in an ancestral population that gave rise to both extant subpopulations. However, it seems clear that a single proto-domesticate population did not give rise to all 4 extant cultivated populations, and at least two separate domestication events are necessary to explain why *temperate japonica* shares no common-by-descent sweep loci with either *indica* and *aus*, and likewise why *aus* does not share common-by-descent sweep loci with either *tropical japonica* and *temperate japonica*. The most likely explanation for why *tropical japonica* shows common sweep loci with both *indica* and *temperate japonica* is that this extant population shares ancestry with both the *proto-indica* and *proto-japonica* early domesticates, but *aus* was differentiated or isolated from *proto-indica* much earlier and the number of selective events inherited into both *aus* and *tropical japonica* from a *proto-indica* ancestor is too small to be significant in the comparison of sHH means.

As discussed above, it is not likely that the classic selective sweep model would capture the specific pattern of diversity produced by introgression where recombinants are being selected for rather than the trait locus alone. Our analysis of haplotype sharing showed extensive level of putative introgression of selected loci originating in a different subpopulation. Future work will seek to identify the target genes underlying these loci and determine with more precise methods whether or not introgression has occurred, where the allele originated, and what favorable traits are conveyed.

An interesting note is that selective sweeps detected only in *tropical japonica* appear to have been introgressed into both *temperate japonica* and *indica*. Conversely, the *aus* subpopulation appears to have neither received a significant amount of introgression of selected loci from other populations, nor donated selected alleles of *aus* origin. Instead, it appears to be more diverged at its selected loci from other subpopulations, suggesting selection in this subpopulation is occurring independently of others.

An on-going debate in rice is whether or not rice was domesticated once, or twice (once in *indica/aus*, again in *temperate/tropical japonica*). Genome-wide  $F_{ST}$  values (not shown) in this study and others clearly show *indica* and *aus* are genetically more similar as are *temperate japonica* and *tropical japonica* and the two groups are quite differentiated from each other. Clearly, domestication is a selection intensive process, and thus identifying and understanding the selective history of these subpopulations has direct implications on how they came to be domesticated, cultivated rice populations. Overall, our results indicate a largely independent selective history of each subpopulation, but extensive introgression of advantageous alleles between subpopulations has likely occurred. While the ongoing debate focuses on whether or not *indica/aus* and *temperate/tropical japonica* are a separate (Caicedo *et al.*, 2007; Guo *et al.*, 2008) or single domestication event (Gao and Innan, 2008; Vaughan *et al.*, 2008) (see also Sweeney and McCouch (2007); Panaud (2009)), our data suggest that *aus* may have a nearly completely independent selective history from any other subpopulation and might be considered a third independent domestication. From the haplotype sharing analysis, there is little or no introgressive contact between *aus* and *indica* but clear support for introgression from *tropical japonica* into *temperate japonica*. However, the number of shared sweeps between *indica* and *aus* was highly significant with an elevation of haplotype sharing at these loci, suggesting that these two subpopulations may share ancestry

with an early domesticated population, but more recent contact has been infrequent and continued domestication in *aus* occurred independently of *indica* and other subpopulations.

While this study aims to provide the most comprehensive analysis of selection across the rice genome to date, we hasten to point out its limitations. The classic selective sweep model of a novel mutation immediately resulting a beneficial phenotype and rising rapidly to fixation may only describe a minority of selective events in the history of cultivated rice. In addition to not describing introgression events well, the classic selective sweep model does not describe selection on standing variation. In the classic selective sweep model, the mutation is either fully dominant phenotypically or the system is haploid and immediately upon occurrence a selective advantage is conveyed. If a mutation is recessive, it must increase in frequency via drift before homozygous genotypes are likely and a phenotypic consequence can be realized. During such time, it may recombine onto several backgrounds resulting in several haplotypes potentially entering the selective phase. Many domestication genes which have been described in rice are loss-of-function recessive mutations. If these mutations occurred in the wild, they may segregate at low frequencies in the population phenotypically hidden in heterozygotes for extended periods of time. Upon inbreeding in cultivation, the recessive phenotype would be revealed and be selected if advantageous for cultivation. Our analysis would have low power to detect such selective events, and early domestication traits may have primarily utilized standing variation in the wild. Thus, our inability to find a common shared selective history of all cultivated rice may be due to lack of power and incorrect model to detect this sort of selective event. Regardless, if domestication is viewed as an ongoing process, each subpopulation clearly has a largely distinct selective history.



Although all attempts have been made to maximize power and obtain accurate statistical significance of the tests of selection performed in this study, and the dataset employed is the most comprehensive genome-wide dataset to date in rice, it is likely that each individual subpopulation remains under-powered with minimal at best sample depths obtained in each. Future studies with more ample sample sizes may discover more selective targets which change the pairwise comparisons and thus the interpretations put forth here. A significant limitation of our methods is that the selective sweep model, inherited from Nielsen *et al.* (2005) and Kim and Stephan (2002) assumes that the fixation of the selected allele occurred just prior to sampling of the data and that no new mutation has occurred. More ancient selective sweeps associated with the first domestication events may have since accumulated mutations and are not detected by these methods. Improved methods in future studies can hopefully incorporate a time parameter to the classic selective sweep model such that the possibility of new mutations, and their frequency relative to the time since fixation, can be accounted for and appropriately modeled, increasing the power to detect such ancient selective fixations.

One aspect of these results we have not analyzed is what genes underlie these predicted selection targets and what phenotypes they might convey. Figure 4.4 shows several recently cloned genes in rice many of which have phenotypes thought to be important in cultivation and domestication and some of which population genetic data has shown evidence for selection (Takahashi *et al.*, 2001; Fan *et al.*, 2006; Konishi *et al.*, 2006; Furukawa *et al.*, 2007; Shomura *et al.*, 2008; Shan *et al.*, 2009; Zhang *et al.*, 2009b,a; Sugimoto *et al.*, 2010). While in many cases a selective target is shown directly on top of one of these loci in at least one subpopulation, or in an adjacent 100 Kb window, the accuracy of target prediction with sample depths used in this study is too crude to pinpoint a single gene. With the high gene density in the rice genome, a

100 Kb window centered on a target prediction may contain dozens of genes with no reason based on the CL analysis to favor any one of them. However, the demonstration of introgression of these selective events into other subpopulations presents an exciting opportunity to pursue further in that identifying and localizing the introgression may further refine the target region such that a single gene or a manageable number of genes may be pursued further in functional studies.

## REFERENCES

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A., Lewis, S. E., Richards, S., Ashburner, M., Henderson, S. N., Sutton, G. G., Wortman, J. R., Yandell, M. D., Zhang, Q., Chen, L. X., Brandon, R. C., Rogers, Y. H., Blazej, R. G., Champe, M., Pfeiffer, B. D., Wan, K. H., Doyle, C., Baxter, E. G., Helt, G., Nelson, C. R., Gabor, G. L., Abril, J. F., Agbayani, A., An, H. J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R. M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E. M., Beeson, K. Y., Benos, P. V., Berman, B. P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M. R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K. C., Busam, D. A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J. M., Cawley, S., Dahlke, C., Davenport, L. B., Davies, P., de Pablos, B., Delcher, A., Deng, Z., Mays, A. D., Dew, I., Dietz, S. M., Dodson, K., Doup, L. E., Downes, M., Dugan-Rocha, S., Dunkov, B. C., Dunn, P., Durbin, K. J., Evangelista, C. C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A. E., Garg, N. S., Gelbart, W. M., Glasser, K., Glodek, A., Gong, F., Gorrell, J. H., Gu, Z., Guan, P., Harris, M., Harris, N. L., Harvey, D., Heiman, T. J., Hernandez, J. R., Houck, J., Hostin, D., Houston, K. A., Howland, T. J., Wei, M. H., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G. H., Ke, Z., Kennison, J. A., Ketchum, K. A., Kimmel, B. E., Kodira, C. D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A. A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T. C., McLeod, M. P., McPherson, D., Merkulov, G., Milshina, N. V., Mobarri, C., Morris, J., Moshrefi, A., Mount, S. M., Moy, M., Murphy, B., Murphy, L., Muzny, D. M., Nelson, D. L., Nelson, D. R., Nelson, K. A., Nixon, K., Nusskern, D. R., Pacleb, J. M., Palazzolo, M., Pittman, G. S., Pan, S., Pollard, J., Puri, V., Reese, M. G., Reinert, K., Remington, K., Saunders, R. D., Scheeler, F., Shen, H., Shue, B. C., Sidn-Kiamos, I., Simpson,

- M., Skupski, M. P., Smith, T., Spier, E., Spradling, A. C., Stapleton, M., Strong, R., Sun, E., Svirskas, R., Tector, C., Turner, R., Venter, E., Wang, A. H., Wang, X., Wang, Z. Y., Wassarman, D. A., Weinstock, G. M., Weissenbach, J., Williams, S. M., WoodageT, Worley, K. C., Wu, D., Yang, S., Yao, Q. A., Ye, J., Yeh, R. F., Zaveri, J. S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X. H., Zhong, F. N., Zhong, W., Zhou, X., Zhu, S., Zhu, X., Smith, H. O., Gibbs, R. A., Myers, E. W., Rubin, G. M., and Venter, J. C. (2000). The genome sequence of drosophila melanogaster. *Science*, **287**(5461), 2185–2195.
- Affymetrix Inc. (2006). BRLMM: an improved genotype calling method for the genechip mapping 500k array set. [http://affymetrix.com/support/technical/whitepapers/brlmm\\_whitepaper.pdf](http://affymetrix.com/support/technical/whitepapers/brlmm_whitepaper.pdf).
- Altschul, S. F., Madden, T. L., Schffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, **25**(17), 3389–3402.
- Antequera, F. and Bird, A. P. (1988). Unmethylated cpg islands associated with genes in higher plant dna. *EMBO J*, **7**(8), 2295–2299.
- Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *Nature*, **408**(6814), 796–815.
- Barbazuk, W. B., Emrich, S. J., Chen, H. D., Li, L., and Schnable, P. S. (2007). Snp discovery via 454 transcriptome sequencing. *Plant J*, **51**(5), 910–918.
- Bennetzen, J. L., Schrick, K., Springer, P. S., Brown, W. E., and SanMiguel, P. (1994). Active maize genes are unmodified and flanked by diverse classes of modified, highly repetitive dna. *Genome*, **37**(4), 565–576.

- Blanc, G. and Wolfe, K. H. (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell*, **16**(7), 1667–1678.
- Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., and Shao, Y. (1997). The complete genome sequence of escherichia coli k-12. *Science*, **277**(5331), 1453–1462.
- Buckler, E. S., Thornsberry, J. M., and Kresovich, S. (2001). Molecular diversity, structure and domestication of grasses. *Genet Res*, **77**(3), 213–218.
- Buckler, E. S., Holland, J. B., Bradbury, P. J., Acharya, C. B., Brown, P. J., Browne, C., Ersoz, E., Flint-Garcia, S., Garcia, A., Glaubitz, J. C., Goodman, M. M., Harjes, C., Guill, K., Kroon, D. E., Larsson, S., Lepak, N. K., Li, H., Mitchell, S. E., Pressoir, G., Peiffer, J. A., Rosas, M. O., Rocheford, T. R., Roday, M. C., Romero, S., Salvo, S., Villeda, H. S., da Silva, H. S., Sun, Q., Tian, F., Upadaya, N., Ware, D., Yates, H., Yu, J., Zhang, Z., Kresovich, S., and McMullen, M. D. (2009). The genetic architecture of maize flowering time. *Science*, **325**(5941), 714–718.
- Caicedo, A. L., Williamson, S. H., Hernandez, R. D., Boyko, A., Fledel-Alon, A., York, T. L., Polato, N. R., Olsen, K. M., Nielsen, R., McCouch, S. R., Bustamante, C. D., and Purugganan, M. D. (2007). Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genet*, **3**(9), 1745–1756.
- Carvalho, B., Bengtsson, H., Speed, T. P., and Irizarry, R. A. (2007). Exploration, normalization, and genotype calls of high-density oligonucleotide snp array data. *Biostatistics*, **8**(2), 485–499.
- Cervera, M., Ruiz-Garcia, L., and Martinez-Zapater, J. (2002). Analysis of DNA methy-

- lation in *Arabidopsis thaliana* based on methylation-sensitive AFLP markers. *Molecular Genetics and Genomics*, **268**(4), 543–552.
- Clark, A. G., Hubisz, M. J., Bustamante, C. D., Williamson, S. H., and Nielsen, R. (2005). Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res*, **15**(11), 1496–1502.
- Consortium, I. H. (2003). The international hapmap project. *Nature*, **426**(6968), 789–796.
- Droege, M. and Hill, B. (2008). The genome sequencer flx system—longer reads, more applications, straight forward bioinformatics and more complete data sets. *J Biotechnol*, **136**(1-2), 3–10.
- Emberton, J., Ma, J., Yuan, Y., SanMiguel, P., and Bennetzen, J. L. (2005). Gene enrichment in maize with hypomethylated partial restriction (hmpr) libraries. *Genome Res*, **15**(10), 1441–1446.
- Emrich, S. J., Li, L., Wen, T.-J., Yandea-Nelson, M. D., Fu, Y., Guo, L., Chou, H.-H., Aluru, S., Ashlock, D. A., and Schnable, P. S. (2007). Nearly identical paralogs: implications for maize (*zea mays* l.) genome evolution. *Genetics*, **175**(1), 429–439.
- Fan, C., Xing, Y., Mao, H., Lu, T., Han, B., Xu, C., Li, X., and Zhang, Q. (2006). Gs3, a major qtl for grain length and weight and minor qtl for grain width and thickness in rice, encodes a putative transmembrane protein. *Theor Appl Genet*, **112**(6), 1164–1171.
- Fan, J. B., Oliphant, A., Shen, R., Kermani, B. G., Garcia, F., Gunderson, K. L., Hansen, M., Steemers, F., Butler, S. L., Deloukas, P., Galver, L., Hunt, S., McBride, C., Bibikova, M., Rubano, T., Chen, J., Wickham, E., Doucet, D., Chang, W., Campbell, D., Zhang, B., Kruglyak, S., Bentley, D., Haas, J., Rigault, P., Zhou, L., Stuelpnagel,

- J., and Chee, M. S. (2003). Highly parallel snp genotyping. *Cold Spring Harb Symp Quant Biol*, **68**, 69–78.
- Fay, J. C. and Wu, C. I. (2000). Hitchhiking under positive darwinian selection. *Genetics*, **155**(3), 1405–1413.
- Finnegan, E. J., Peacock, W. J., and Dennis, E. S. (2000). Dna methylation, a key regulator of plant development and other processes. *Curr Opin Genet Dev*, **10**(2), 217–223.
- Fu, H., Park, W., Yan, X., Zheng, Z., Shen, B., and Dooner, H. K. (2001). The highly recombinogenic bz locus lies in an unusually gene-rich region of the maize genome. *Proc Natl Acad Sci U S A*, **98**(15), 8903–8908.
- Fu, H., Zheng, Z., and Dooner, H. K. (2002). Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. *Proc Natl Acad Sci U S A*, **99**(2), 1082–1087.
- Fu, Y. X. and Li, W. H. (1993). Statistical tests of neutrality of mutations. *Genetics*, **133**(3), 693–709.
- Furukawa, T., Maekawa, M., Oki, T., Suda, I., Iida, S., Shimada, H., Takamure, I., and ichi Kadowaki, K. (2007). The rc and rd genes are involved in proanthocyanidin synthesis in rice pericarp. *Plant J*, **49**(1), 91–102.
- Gao, L.-Z. and Innan, H. (2008). Nonindependent domestication of the two rice subspecies, *oryza sativa* ssp. *indica* and ssp. *japonica*, demonstrated by multilocus microsatellites. *Genetics*, **179**(2), 965–976.
- Garris, A. J., Tai, T. H., Coburn, J., Kresovich, S., and McCouch, S. (2005). Genetic structure and diversity in *oryza sativa* l. *Genetics*, **169**(3), 1631–1638.

- Gaut, B. S. and Doebley, J. F. (1997). Dna sequence evidence for the segmental allotetraploid origin of maize. *Proc Natl Acad Sci U S A*, **94**(13), 6809–6814.
- Goff, S. A., Ricke, D., Lan, T.-H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., Hadley, D., Hutchison, D., Martin, C., Katagiri, F., Lange, B. M., Moughamer, T., Xia, Y., Budworth, P., Zhong, J., Miguel, T., Paszkowski, U., Zhang, S., Colbert, M., Lin Sun, W., Chen, L., Cooper, B., Park, S., Wood, T. C., Mao, L., Quail, P., Wing, R., Dean, R., Yu, Y., Zharkikh, A., Shen, R., Sahasrabudhe, S., Thomas, A., Cannings, R., Gutin, A., Pruss, D., Reid, J., Tavtigian, S., Mitchell, J., Eldredge, G., Scholl, T., Miller, R. M., Bhatnagar, S., Adey, N., Rubano, T., Tusneem, N., Robinson, R., Feldhaus, J., Macalma, T., Oliphant, A., and Briggs, S. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science*, **296**(5565), 92–100.
- Gore, M., Bradbury, P., Hogers, R., Kirst, M., Verstege, E., van Oeveren, J., Peleman, J., Buckler, E., and van Eijk, M. (2007). Evaluation of target preparation methods for single-feature polymorphism detection in large complex plant genomes. *Crop Science*, **47**(S2).
- Guo, X., Ruan, S., Hu, W., Cai, D., and Fan, L. (2008). Chloroplast dna insertions into the nuclear genome of rice: the genes, sites and ages of insertion involved. *Funct Integr Genomics*, **8**(2), 101–108.
- Hake, S. and Walbot, V. (1980). The genome of *Zea mays*, its organization and homology to related grasses. *Chromosoma*, **79**(3), 251–270.
- Haldane, J. B. S. (1945). On a method of estimating frequencies. *Biometrika*, **33**(3), 222–225.



- Henry, A. M. and Damerval, C. (1997). High rates of polymorphism and recombination at the opaque-2 locus in cultivated maize. *Mol Gen Genet*, **256**(2), 147–157.
- Hirschhorn, J. N. and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet*, **6**(2), 95–108.
- Huang, X. and Madan, A. (1999). Cap3: A dna sequence assembly program. *Genome Res*, **9**(9), 868–877.
- Hudson, R. R. (2002). Generating samples under a wright-fisher neutral model of genetic variation. *Bioinformatics*, **18**(2), 337–338.
- Jensen, J. D., Kim, Y., DuMont, V. B., Aquadro, C. F., and Bustamante, C. D. (2005). Distinguishing between selective sweeps and demography using dna polymorphism data. *Genetics*, **170**(3), 1401–1410.
- Kalyanaraman, A., Emrich, S., Schnable, P., and Aluru, S. (2007). Assembling genomes on large-scale parallel computers. *Journal of Parallel and Distributed Computing*, **67**(12), 1240–1255.
- Kent, W. J. (2002). Blat—the blast-like alignment tool. *Genome Res*, **12**(4), 656–664.
- Kim, S. and Misra, A. (2007). Snp genotyping: technologies and biomedical applications. *Annu Rev Biomed Eng*, **9**, 289–320.
- Kim, Y. and Stephan, W. (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*, **160**(2), 765–777.
- Konishi, S., Izawa, T., Lin, S. Y., Ebana, K., Fukuta, Y., Sasaki, T., and Yano, M. (2006). An snp caused loss of seed shattering during rice domestication. *Science*, **312**(5778), 1392–1396.

- Korn, J. M., Kuruvilla, F. G., McCarroll, S. A., Wysoker, A., Nemesh, J., Cawley, S., Hubbell, E., Veitch, J., Collins, P. J., Darvishi, K., Lee, C., Nizzari, M. M., Gabriel, S. B., Purcell, S., Daly, M. J., and Altshuler, D. (2008). Integrated genotype calling and association analysis of snps, common copy number polymorphisms and rare cnvs. *Nat Genet*, **40**(10), 1253–1260.
- Lai, J., Ma, J., Swigonov, Z., Ramakrishna, W., Linton, E., Llaca, V., Tanyolac, B., Park, Y.-J., Jeong, O.-Y., Bennetzen, J. L., and Messing, J. (2004). Gene loss and movement in the maize genome. *Genome Res*, **14**(10A), 1924–1931.
- Lander, E., Linton, L., Birren, B., Nusbaum, C., Zody, M., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**(6822), 860–921.
- Lander, E. S. and Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, **2**(3), 231–239.
- Lee, M., Sharopova, N., Beavis, W. D., Grant, D., Katt, M., Blair, D., and Hallauer, A. (2002). Expanding the genetic map of maize with the intermated b73 x mo17 (ibm) population. *Plant Mol Biol*, **48**(5-6), 453–461.
- Li, W., Ruf, S., and Bock, R. (2006). Constancy of organellar genome copy numbers during leaf development and senescence in higher plants. *Mol Genet Genomics*, **275**(2), 185–192.
- Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., and Ecker, J. R. (2008). Highly integrated single-base resolution maps of the epigenome in arabidopsis. *Cell*, **133**(3), 523–536.
- Liu, W., Di, X., Yang, G., Matsuzaki, H., Huang, J., Mei, R., Ryder, T. B., Webster,

- T. A., Dong, S., Liu, G., Jones, K. W., Kennedy, G. C., and Kulp, D. (2003). Algorithms for large-scale genotyping microarrays. *Bioinformatics*, **19**(18), 2397–2403.
- Lynch, M., Sung, W., Morris, K., Coffey, N., Landry, C. R., Dopman, E. B., Dickinson, W. J., Okamoto, K., Kulkarni, S., Hartl, D. L., and Thomas, W. K. (2008). A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc Natl Acad Sci U S A*, **105**(27), 9272–9277.
- Marezzo, K. and Broeckel, U. (2008). Genotyping platforms for mass-throughput genotyping with snps, including human genome-wide scans. *Adv Genet*, **60**, 107–139.
- Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., and Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**(7057), 376–380.
- Marth, G. T., Korf, I., Yandell, M. D., Yeh, R. T., Gu, Z., Zakeri, H., Stitzel, N. O., Hillier, L., Kwok, P. Y., and Gish, W. R. (1999). A general approach to single-nucleotide polymorphism discovery. *Nat Genet*, **23**(4), 452–456.
- Matsuzaki, H., Dong, S., Loi, H., Di, X., Liu, G., Hubbell, E., Law, J., Berntsen, T., Chadha, M., Hui, H., Yang, G., Kennedy, G. C., Webster, T. A., Cawley, S., Walsh,

- P. S., Jones, K. W., Fodor, S. P. A., and Mei, R. (2004). Genotyping over 100,000 snps on a pair of oligonucleotide arrays. *Nat Methods*, **1**(2), 109–111.
- McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet*, **9**(5), 356–369.
- McNally, K. L., Childs, K. L., Bohnert, R., Davidson, R. M., Zhao, K., Ulat, V. J., Zeller, G., Clark, R. M., Hoen, D. R., Bureau, T. E., Stokowski, R., Ballinger, D. G., Frazer, K. A., Cox, D. R., Padhukasahasram, B., Bustamante, C. D., Weigel, D., Mackill, D. J., Bruskiewich, R. M., Rtsch, G., Buell, C. R., Leung, H., and Leach, J. E. (2009). Genomewide snp variation reveals relationships among landraces and modern varieties of rice. *Proc Natl Acad Sci U S A*, **106**(30), 12273–12278.
- Meselson, M. and Yuan, R. (1968). Dna restriction enzyme from e. coli. *Nature*, **217**(5134), 1110–1114.
- Messing, J., Bharti, A. K., Karlowski, W. M., Gundlach, H., Kim, H. R., Yu, Y., Wei, F., Fuks, G., Soderlund, C. A., Mayer, K. F. X., and Wing, R. A. (2004). Sequence composition and genome organization of maize. *Proc Natl Acad Sci U S A*, **101**(40), 14349–14354.
- Meyers, B. C., Tingey, S. V., and Morgante, M. (2001). Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. *Genome Res*, **11**(10), 1660–1676.
- Morgan, T., Sturtevant, A., H.J, H. M., and Bridges, C. (1915). *The mechanism of Mendelian heredity*. Holt Rinehart & Winston, New York.
- Mullis, K. and Faloona, F. (1987). Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods in enzymology*, **155**, 335.

- Nielsen, R. (2004). Population genetic analysis of ascertained snp data. *Hum Genomics*, **1**(3), 218–224.
- Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., and Bustamante, C. (2005). Genomic scans for selective sweeps using snp data. *Genome Res*, **15**(11), 1566–1575.
- Palmer, L. E., Rabinowicz, P. D., O’Shaughnessy, A. L., Balija, V. S., Nascimento, L. U., Dike, S., de la Bastide, M., Martienssen, R. A., and McCombie, W. R. (2003). Maize genome sequencing by methylation filtration. *Science*, **302**(5653), 2115–2117.
- Panaud, O. (2009). The molecular bases of cereal domestication and the history of rice. *C R Biol*, **332**(2-3), 267–272.
- Paterson, A. H., Lander, E. S., Hewitt, J. D., Peterson, S., Lincoln, S. E., and Tanksley, S. D. (1988). Resolution of quantitative traits into mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature*, **335**(6192), 721–726.
- Possingham, J. (1980). Plastid replication and development in the life cycle of higher plants. *Annual Review of Plant Physiology*, **31**(1), 113–129.
- Przeworski, M. (2002). The signature of positive selection at randomly chosen loci. *Genetics*, **160**(3), 1179–1189.
- Rabbee, N. and Speed, T. P. (2006). A genotype calling algorithm for affymetrix snp arrays. *Bioinformatics*, **22**(1), 7–12.
- Rabinowicz, P. D. (2003). Constructing gene-enriched plant genomic libraries using methylation filtration technology. *Methods Mol Biol*, **236**, 21–36.

- Rabinowicz, P. D., Schutz, K., Dedhia, N., Yordan, C., Parnell, L. D., Stein, L., McCombie, W. R., and Martienssen, R. A. (1999). Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. *Nat Genet*, **23**(3), 305–308.
- Rabinowicz, P. D., Citek, R., Budiman, M. A., Nunberg, A., Bedell, J. A., Lakey, N., O’Shaughnessy, A. L., Nascimento, L. U., McCombie, W. R., and Martienssen, R. A. (2005). Differential methylation of genes and repeats in land plants. *Genome Res*, **15**(10), 1431–1440.
- Remington, D. L., Thornsberry, J. M., Matsuoka, Y., Wilson, L. M., Whitt, S. R., Doebley, J., Kresovich, S., Goodman, M. M., and Buckler, E. S. (2001). Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc Natl Acad Sci U S A*, **98**(20), 11479–11484.
- Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., and Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **419**(6909), 832–837.
- SanMiguel, P., Tikhonov, A., Jin, Y. K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P. S., Edwards, K. J., Lee, M., Avramova, Z., and Bennetzen, J. L. (1996). Nested retrotransposons in the intergenic regions of the maize genome. *Science*, **274**(5288), 765–768.
- Shan, J.-X., Zhu, M.-Z., Shi, M., Gao, J.-P., and Lin, H.-X. (2009). Fine mapping and candidate gene analysis of *spd6*, responsible for small panicle and dwarfness in wild rice (*Oryza rufipogon* Griff.). *Theor Appl Genet*, **119**(5), 827–836.

- Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D., and Church, G. M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, **309**(5741), 1728–1732.
- Shomura, A., Izawa, T., Ebana, K., Ebitani, T., Kanegae, H., Konishi, S., and Yano, M. (2008). Deletion in a gene associated with grain size increased yields during rice domestication. *Nat Genet*, **40**(8), 1023–1028.
- Smith, D. R., Quinlan, A. R., Peckham, H. E., Makowsky, K., Tao, W., Woolf, B., Shen, L., Donahue, W. F., Tusneem, N., Stromberg, M. P., Stewart, D. A., Zhang, L., Ranade, S. S., Warner, J. B., Lee, C. C., Coleman, B. E., Zhang, Z., McLaughlin, S. F., Malek, J. A., Sorenson, J. M., Blanchard, A. P., Chapman, J., Hillman, D., Chen, F., Rokhsar, D. S., McKernan, K. J., Jeffries, T. W., Marth, G. T., and Richardson, P. M. (2008). Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Res*, **18**(10), 1638–1642.
- Smith, J. M. and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genet Res*, **23**(1), 23–35.
- Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol*, **147**(1), 195–197.
- Sturtevant, A. (1913). The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. *JOURNAL OF EXPERIMENTAL ZOOLOGY*, **14**(1), 43–59.
- Sugimoto, K., Takeuchi, Y., Ebana, K., Miyao, A., Hirochika, H., Hara, N., Ishiyama, K., Kobayashi, M., Ban, Y., Hattori, T., and Yano, M. (2010). Molecular cloning

- of *sdr4*, a regulator involved in seed dormancy and domestication of rice. *Proc Natl Acad Sci U S A*, **107**(13), 5792–5797.
- Sweeney, M. and McCouch, S. (2007). The complex history of the domestication of rice. *Ann Bot*, **100**(5), 951–957.
- Swigonova, Z., Lai, J., Ma, J., Ramakrishna, W., Llaca, V., Bennetzen, J. L., and Messing, J. (2004). Close split of sorghum and maize genome progenitors. *Genome Res*, **14**(10A), 1916–1923.
- Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics*, **123**(3), 585–595.
- Takahashi, Y., Shomura, A., Sasaki, T., and Yano, M. (2001). Hd6, a rice quantitative trait locus involved in photoperiod sensitivity, encodes the alpha subunit of protein kinase ck2. *Proc Natl Acad Sci U S A*, **98**(14), 7922–7927.
- Tenaillon, M. I., Sawkins, M. C., Long, A. D., Gaut, R. L., Doebley, J. F., and Gaut, B. S. (2001). Patterns of dna sequence polymorphism along chromosome 1 of maize (*zea mays* ssp. *mays* l.). *Proc Natl Acad Sci U S A*, **98**(16), 9161–9166.
- Teo, Y. Y., Inouye, M., Small, K. S., Gwilliam, R., Deloukas, P., Kwiatkowski, D. P., and Clark, T. G. (2007). A genotype calling algorithm for the illumina beadarray platform. *Bioinformatics*, **23**(20), 2741–2746.
- The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, **437**(7063), 1299–1320.
- The International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million snps. *Nature*, **449**(7164), 851–861.



- Vaughan, D., Lu, B., and Tomooka, N. (2008). The evolving story of rice evolution. *Plant Science*, **174**(4), 394–408.
- Vaughn, M. W., Tanurdzi, M., Lippman, Z., Jiang, H., Carrasquillo, R., Rabinowicz, P. D., Dedhia, N., McCombie, W. R., Agier, N., Bulski, A., Colot, V., Doerge, R. W., and Martienssen, R. A. (2007). Epigenetic natural variation in *arabidopsis thaliana*. *PLoS Biol*, **5**(7), e174.
- Venter, J., Adams, M., Myers, E., Li, P., Mural, R., Sutton, G., Smith, H., Yandell, M., Evans, C., Holt, R., *et al.* (2001). The sequence of the human genome. *science*, **291**(5507), 1304.
- Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biol*, **4**(3), e72.
- Wakeley, J. and Aliacar, N. (2001). Gene genealogies in a metapopulation. *Genetics*, **159**(2), 893–905.
- Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, **171**(4356), 737–738.
- Wei, F., Coe, E., Nelson, W., Bharti, A. K., Engler, F., Butler, E., Kim, H., Goicoechea, J. L., Chen, M., Lee, S., Fuks, G., Sanchez-Villeda, H., Schroeder, S., Fang, Z., McMullen, M., Davis, G., Bowers, J. E., Paterson, A. H., Schaeffer, M., Gardiner, J., Cone, K., Messing, J., Soderlund, C., and Wing, R. A. (2007). Physical and genetic structure of the maize genome reflects its complex evolutionary history. *PLoS Genet*, **3**(7), e123.
- Whitelaw, C. A., Barbazuk, W. B., Perte, G., Chan, A. P., Cheung, F., Lee, Y., Zheng, L., van Heeringen, S., Karamycheva, S., Bennetzen, J. L., SanMiguel, P., Lakey, N.,

- Bedell, J., Yuan, Y., Budiman, M. A., Resnick, A., Aken, S. V., Utterback, T., Riedmuller, S., Williams, M., Feldblyum, T., Schubert, K., Beachy, R., Fraser, C. M., and Quackenbush, J. (2003). Enrichment of gene-coding sequences in maize by genome filtration. *Science*, **302**(5653), 2118–2120.
- Williamson, S. H., Hubisz, M. J., Clark, A. G., Payseur, B. A., Bustamante, C. D., and Nielsen, R. (2007). Localizing recent adaptive evolution in the human genome. *PLoS Genet*, **3**(6), e90.
- Wright, S. I., Bi, I. V., Schroeder, S. G., Yamasaki, M., Doebley, J. F., McMullen, M. D., and Gaut, B. S. (2005). The effects of artificial selection on the maize genome. *Science*, **308**(5726), 1310–1314.
- Yamasaki, M., Tenaillon, M. I., Bi, I. V., Schroeder, S. G., Sanchez-Villeda, H., Doebley, J. F., Gaut, B. S., and McMullen, M. D. (2005). A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. *Plant Cell*, **17**(11), 2859–2872.
- Yang, H., Ding, Y., Hutchins, L. N., Szatkiewicz, J., Bell, T. A., Paigen, B. J., Graber, J. H., de Villena, F. P.-M., and Churchill, G. A. (2009). A customized and versatile high-density genotyping array for the mouse. *Nat Methods*, **6**(9), 663–666.
- Yao, H., Zhou, Q., Li, J., Smith, H., Yandea, M., Nikolau, B. J., and Schnable, P. S. (2002). Molecular characterization of meiotic recombination across the 140-kb multi-genic a1-sh2 interval of maize. *Proc Natl Acad Sci U S A*, **99**(9), 6157–6162.
- Yu, J., Hu, S., Wang, J., Wong, G. K.-S., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., Cao, M., Liu, J., Sun, J., Tang, J., Chen, Y., Huang, X., Lin, W., Ye, C., Tong, W., Cong, L., Geng, J., Han, Y., Li, L., Li, W., Hu, G., Huang, X., Li, W., Li, J., Liu, Z., Li, L., Liu, J., Qi, Q., Liu, J., Li, L., Li, T., Wang, X., Lu, H., Wu, T., Zhu,

- M., Ni, P., Han, H., Dong, W., Ren, X., Feng, X., Cui, P., Li, X., Wang, H., Xu, X., Zhai, W., Xu, Z., Zhang, J., He, S., Zhang, J., Xu, J., Zhang, K., Zheng, X., Dong, J., Zeng, W., Tao, L., Ye, J., Tan, J., Ren, X., Chen, X., He, J., Liu, D., Tian, W., Tian, C., Xia, H., Bao, Q., Li, G., Gao, H., Cao, T., Wang, J., Zhao, W., Li, P., Chen, W., Wang, X., Zhang, Y., Hu, J., Wang, J., Liu, S., Yang, J., Zhang, G., Xiong, Y., Li, Z., Mao, L., Zhou, C., Zhu, Z., Chen, R., Hao, B., Zheng, W., Chen, S., Guo, W., Li, G., Liu, S., Tao, M., Wang, J., Zhu, L., Yuan, L., and Yang, H. (2002). A draft sequence of the rice genome (*oryza sativa* l. ssp. *indica*). *Science*, **296**(5565), 79–92.
- Yu, J., Holland, J. B., McMullen, M. D., and Buckler, E. S. (2008). Genetic design and statistical power of nested association mapping in maize. *Genetics*, **178**(1), 539–551.
- Yuan, Y., SanMiguel, P. J., and Bennetzen, J. L. (2002). Methylation-spanning linker libraries link gene-rich regions and identify epigenetic boundaries in *zea mays*. *Genome Res*, **12**(9), 1345–1349.
- Yuan, Y., SanMiguel, P. J., and Bennetzen, J. L. (2003). High-cot sequence analysis of the maize genome. *Plant J*, **34**(2), 249–255.
- Zagulski, M., Herbert, C. J., and Rytka, J. (1998). Sequencing and functional analysis of the yeast genome. *Acta Biochim Pol*, **45**(3), 627–643.
- Zhang, L.-B., Zhu, Q., Wu, Z.-Q., Ross-Ibarra, J., Gaut, B. S., Ge, S., and Sang, T. (2009a). Selection on grain shattering genes and rates of rice domestication. *New Phytol*, **184**(3), 708–720.
- Zhang, Y., Wang, J., Zhang, X., Chen, J.-Q., Tian, D., and Yang, S. (2009b). Genetic signature of rice domestication shown by a variety of genes. *J Mol Evol*, **68**(4), 393–402.

- Zhao, W., Canaran, P., Jurkuta, R., Fulton, T., Glaubitz, J., Buckler, E., Doebley, J., Gaut, B., Goodman, M., Holland, J., Kresovich, S., McMullen, M., Stein, L., and Ware, D. (2006). Panzea: a database and resource for molecular and functional diversity in the maize genome. *Nucleic Acids Res*, **34**(Database issue), D752–D757.
- Zhu, C., Gore, M., Buckler, E., and Yu, J. (2008). Status and prospects of association mapping in plants. *The Plant Genome*, **1**(1), 5.